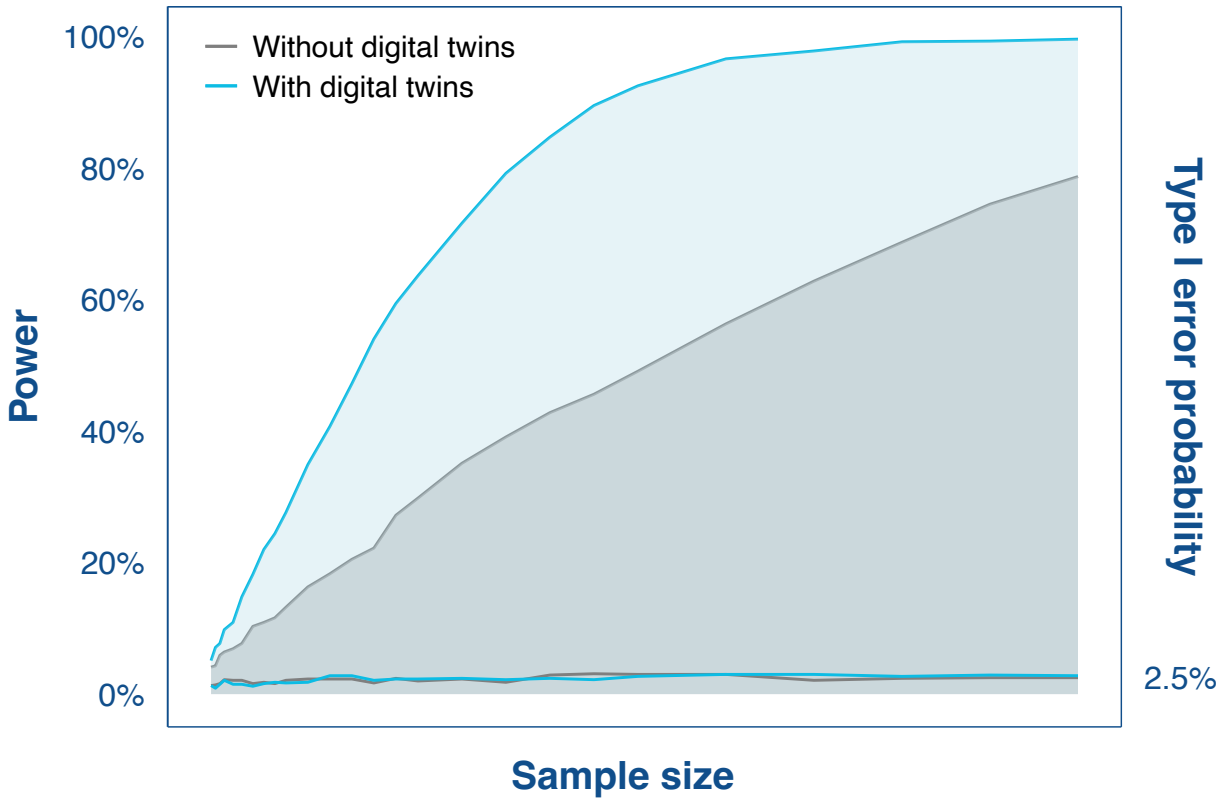
Increasing the power in randomised clinical trials using digital twins



Master Thesis

Emilie Højbjerre-Frandsen
Mathias Lerbech Jeppesen
Rasmus Kuhr Jensen

June 2022

AALBORG UNIVERSITY
STUDENT REPORT

Department of Mathematical Sciences
Aalborg University
Skjernvej 4A
9220 Aalborg Øst
Denmark
http://math.aau.dk

Title:

Increasing the power in randomised clinical trials using digital twins

Project:

3rd and 4th semester master project

Project Period:

September 2021 – June 2022

Participants:

Emilie Højbjerre-Frandsen
Mathias Lerbech Jeppesen
Rasmus Kuhr Jensen

Supervisor:

Mikkel Meyer Andersen

Co-supervisor:

Steffen Falgreen Larsen

Pages: 151

Finished: 2nd of June 2022

Abstract:

This thesis is a study of approaches to leveraging historical data for increasing power in randomised clinical trials (RCTs) with a continuously measured efficacy outcome. Existing methods based on populating the control arm with a synthetic control arm (SCA) fail to strictly control the type I error rate. Therefore, we focus on the novel statistical method of using digital twins (artificially generated patients receiving control medication) in an analysis of covariance (ANCOVA) model.

We show analytically that under certain assumptions, by adjusting for the predicted outcome of a digital twin in an ANCOVA model, we obtain asymptotic efficiency of the average treatment effect estimator among a large class of estimators. This efficiency gain can then be used to decrease the sample size needed in a trial, while maintaining the same power and strictly controlling the type I error rate.

In a simulation study, we compare the performance of an existing SCA approach with the performance of the novel digital twins approach in terms of power gain and type I error rate control. Under several scenarios, we find that the SCA approach provides at best only modest gains in power and is unreliable in terms of controlling the type I error rate. Conversely, the digital twins approach provides strict type I error control and a substantial increase in power, even when assumptions on analytical results are violated.

Lastly, we evaluate the method of digital twins in real world data originating from RCTs previously conducted at Novo Nordisk A/S. We find that the method manages to decrease the required number of subjects in a trial from 83 to 72, with possible improvements by further fine-tuning the method.

Preface

This thesis has been written throughout the course of he 3rd and 4th semesters of the Master's degree programme at the Department of Mathematical Sciences at Aalborg University.

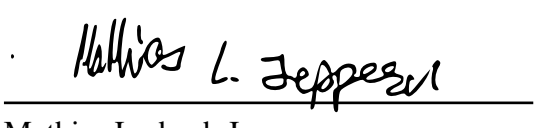
During the process of writing this thesis, some difficulties arose in understanding results from the literature. In this regard, we would like to thank professor William F. Rosenberger and postdoc Alejandro Schuler for taking the time to provide helpful answers to our inquiries.

Additionally, we would like to thank Unlearn.AI for providing insight into using digital twins as an approach in the developing field of leveraging historical data to improve power in clinical studies.

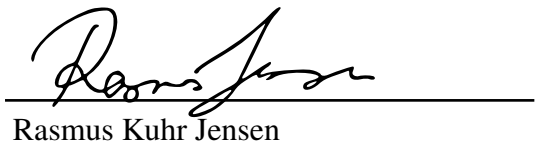
Finally, we would like to express our gratitude to associate professor Mikkel Meyer Andersen and statistical specialist Steffen Falgreen Larsen for supervision during the course of the project period and to Novo Nordisk A/S for providing data for the analysis presented in chapter 7 of this project report.

Signatures

Emilie Højbjerre-Frandsen



Mathias Lerbech Jeppesen



Rasmus Kuhr Jensen

Contents

Preface **i**

1 Introduction **1**
1.1 Reader's Guide . . . 2

2 Randomised Clinical Trials **5**
2.1 Notation and Framework . . . 6
2.2 Randomisation . . . 8
2.2.1 Complete and Forced Balance Randomisation . . . 9
2.2.2 Stratified Randomisation . . . 14
2.3 AN(C)OVA in Randomised Clinical Trials . . . 15
2.3.1 Benefits of Covariate Adjustment in RCTs . . . 17
2.3.2 Regulatory Guidelines for Covariate Adjustment . . . 22

3 Sample Size Calculations **23**
3.1 General Considerations . . . 23
3.2 ANOVA Model z-test . . . 25
3.3 ANOVA Model t-test . . . 30
3.3.1 Approximation Formulas . . . 32
3.4 ANCOVA Model t-test . . . 33
3.4.1 Univariable Covariate Adjustment . . . 33
3.4.2 Approximation Formulas . . . 36
3.4.3 Multivariable Covariate Adjustment . . . 37
3.4.4 Approximation Formulas . . . 39
3.5 Violations of AN(C)OVA Model Assumptions . . . 39
3.6 Multiple Hypotheses Testing . . . 42

4 Synthetic Control Arms **47**
4.1 Propensity Score Matching . . . 48
4.1.1 Estimating the ATT . . . 51
4.1.2 Propensity Score Matching in Clinical Trials . . . 54
4.2 Balance Diagnostics and Variable Selection . . . 58

5 Digital Twins **61**
5.1 The Digital Twins Approach . . . 61
5.2 (Efficient) Influence Functions . . . 63
5.2.1 Efficient Influence Functions of the ATE and Population Mean . . . 65
5.3 Theoretical Properties of the Digital Twins Approach . . . 71
5.3.1 Asymptotic Distributions of AN(C)OVA Estimators . . . 73

5.3.2 Oracle Estimators . . . 85
5.3.3 Adjustment using a Prognostic Model . . . 88
5.4 Sample Size Calculations using Digital Twins . . . 97

6 Simulation Study – Comparison of Approaches 101
6.1 Methods . . . 101
6.1.1 Distribution of Simulated Data . . . 102
6.1.2 AN(C)OVA Models for ATE Estimation . . . 103
6.1.3 Models for Prognostic Score and Propensity Score . . . 104
6.1.4 Performance Measures . . . 106
6.1.5 Prospective Power Estimation . . . 107
6.2 Performance in Different Scenarios . . . 108
6.3 Overspecification and Underspecification . . . 113
6.4 Varying Sample Sizes . . . 117
6.4.1 Performance Assessment . . . 118
6.4.2 Prospective Estimation of Power from Sample Size . . . 122

7 Case Study – Digital Twins in Clinical Trials for Type 2 Diabetes 125
7.1 Data Sets Provided by Novo Nordisk A/S . . . 125
7.1.1 Trial NN1218-3853 . . . 126
7.1.2 Trial NN1218-4049 . . . 127
7.1.3 Trial NN1250-3998 . . . 128
7.2 Curation of Data Sets . . . 129
7.2.1 Standardising Trial-specific Data Sets . . . 129
7.2.2 Current and Historical Data Sets . . . 132
7.3 Methods . . . 135
7.3.1 Variance Estimation . . . 135
7.3.2 The Prognostic Model . . . 135
7.3.3 Post hoc Power Estimation . . . 136
7.4 Results . . . 137

8 Future Perspectives on the Digital Twins Approach 141
8.1 Regulatory Considerations . . . 141
8.2 Further Developments . . . 144
8.2.1 Prognostic Models . . . 144
8.2.2 Bayesian Approaches to using Historical Data . . . 145
8.2.3 Future Research . . . 147

9 Conclusion 149

10 Bibliography 153

Appendices **159**

A AN(C)OVA Model Derivations **161**
A.1 ANOVA Model Maximum Likelihood Estimate . . . 161
A.2 ANOVA Model Variance Estimate . . . 162
A.3 ANCOVA Model Maximum Likelihood Estimate . . . 162
A.4 (Estimated) Variance of the ANCOVA ATE Estimator . . . 165
A.5 Positive Semidefiniteness of Covariance Matrix Difference . . . 167

B FWER Bounds for Multiple Testing Procedures **169**
B.1 Fixed Testing Sequence . . . 169
B.2 Bonferroni Corrections . . . 170

C Theoretical Properties of Digital Twins **173**
C.1 Efficient Influence Functions . . . 173
C.1.1 Lemma 5.2.3 . . . 173
C.2 Theorem 5.3.3 . . . 175
C.2.1 Inverse Matrix of $\mathbb{E}[D^\top D]$ . . . 175
C.2.2 True parameters . . . 177
C.3 Theorem 5.3.4 . . . 178
C.4 Corollary 5.3.5 . . . 182
C.5 Lemma 5.3.7 . . . 184
C.6 Corollary 5.3.8 . . . 186
C.7 Lemma 5.3.10 . . . 187

D Comparison of Approaches **191**
D.1 Performance in Different Scenarios . . . 191
D.2 Overspecification and Underspecification . . . 191
D.3 Varying Sample Sizes . . . 195
D.4 Prospective Power Estimation in Homogeneous Case . . . 197

E Novo Nordisk A/S Clinical Trial Data **199**
E.1 Trial NN1218-3853 . . . 199
E.2 Trial NN1218-4049 . . . 199
E.3 Trial NN1250-3998 . . . 199
E.4 Variables and their Distributions . . . 199

1 Introduction

Randomised clinical trials (RCTs) are conducted with the intent to demonstrate efficacy and safety of a new drug. In this thesis, we will consider RCTs that measure efficacy by the change in a continuously measured outcome after being exposed to the new drug. The participants of RCTs are randomly assigned to either the new drug or a control group receiving placebo or a standard of care drug, and the difference between the outcomes of the two groups is then used to estimate a treatment effect. By this randomised allocation procedure, it is ensured that observed changes in the outcome of the participants can be ascribed to the new drug and are not in fact caused by confounding factors.

Conducting a clinical trial often requires a large number of participants. However, many reasons exist for wanting to minimise the required number of participants needed to be recruited for clinical trials. A large number of participants comes with large economic costs and a long timeline of the trial. Enabling faster clinical trials at a lower cost would provide essential medical treatment to patients in need. In some specific areas, such as pediatric clinical trials, or trials involving rare diseases, recruitment of large groups of participants even constitute a major challenge in the first place. Other considerations such as ethical issues may also be present when large groups of patients are required to be part of the control group of a clinical trial in order to provide certainty when assessing the efficacy of a new drug.

Since medical organisations often possess large amounts of data from the control groups of past clinical trials, one way to increase the efficiency of current trials would be to leverage this data in the current trial using statistical methods. Such methods should ensure large statistical power in regard to detecting a relevant effect size and control the type I error probability such that a non-effective drug is not approved. In this thesis, we investigate the effect of using the novel approach of digital twins in regard to increasing study power in clinical trials by decreasing variability of the estimated treatment effect. Furthermore, we compare this approach to an existing method based on the propensity score, which aims to leverage historical data to increase power in randomised clinical trials.

Within observational studies, methods based on the propensity score provide a well established approach that allows for comparing groups of observations that are not otherwise directly comparable in regard to observed confounding factors. A method using propensity score matching (PSM) has been proposed for the situation in which data from multiple sources are available. This method was recently adapted to the context of RCTs, suggesting to construct an external control arm (SCA) from historical data patients, working as a supplement to an already present control group. In this way, the method seeks to eliminate the effect of confounders while increasing the number of participants, yielding a larger power so that fewer participants need to be recruited. However, this SCA approach comes with no guarantee of strictly controlling the type I

error probability. In addition, the method relies on using only information from the small subset of historical control arm participants that are matched to the patients in the current treatment arm. For these reasons, we find it relevant to investigate the potential gains of employing novel statistical methods that strictly control the type I error probability while using as much information as possible from historical data to increase power in current clinical trials.

A digital twin is an artificial patient not receiving the new drug. This artificial patient has a clinical record generated by a machine learning model trained on historical data. For each participant in the current RCT, a clinical record of a digital twin having the same baseline characteristics is generated. The clinical record of each digital twin contains a predicted outcome of the participant in the (hypothetical) scenario that they received the control treatment. Theoretically, using an analysis of covariace (ANCOVA) model to adjust for the predicted outcomes of these digital twins, the asymptotic variance of the estimated treatment effect is decreased, thereby increasing the power. Thus fewer patients are needed for the clinical trial, thereby enabling smaller and hence more efficient trials.

1.1 Reader's Guide

Chapters 2–4 provide relevant theoretical aspects of randomised clinical trials, sample size calculations and an existing PSM based SCA method for leveraging historical data. Chapters 5–8 describe and discuss analytical properties of the novel approach of digital twins for leveraging historical data as well as investigating its empirical performance in simulated and real world data.

We begin with chapter 2 by introducing the framework and notation of the thesis, and describing some common statistically desirable characteristics of randomised clinical trials. We discuss several types of randomisation, and we show that when we randomise participants, the known and unknown confounders will be evenly distributed between the treatment and control arms. Throughout the thesis, we will be using different specifications of the AN(C)OVA model to estimate the efficacy of a drug in terms of the average treatment effect. Therefore, the chapter also contains a description of how the AN(C)OVA model can be used to obtain such an estimate. In this context, we discuss the benefits of adjusting for covariates, while introducing some of the regulatory guidelines for covariate adjustment.

We continue in chapter 3 by outlining the most common considerations when conducting a sample size calculation in clinical trials. Specifically, we derive formulas for both ANOVA and ANCOVA models for determining the required sample size when a desired power is chosen at a specific significance level. The chapter provides the formal basis for the potential reduction in required sample size obtainable by increasing the number of participants in the trial (which is the method suggested by the SCA approach) and prognostic covariate adjustment (which is the method suggested by the digital twin approach). Furthermore, we discuss how vulnerable the ANCOVA model is in regard to estimating the treatment effect, type I error probability, and power when the model assumptions are violated within RCTs. The chapter ends with briefly considering how to adjust the sample size in the case of multiple endpoints of the trial, which are

considerations that often need to be taken in practice. However, we limit our primary focus of the thesis to considering the situation of a single outcome with a single hypothesis to investigate one endpoint.

The SCA approach is described in chapter 4, where we consider how propensity score matching can be used both in the contexts of the current trial including only a treatment arm and when a control arm is present as well. We begin by describing some specific assumptions for the historical data in order for this to be well suited for use in a current trial. We describe how propensity score matching can be theoretically used to match historical patients to patients in the current treatment arm based on known confounders. Additionally, we describe how the method can be adjusted to the context of two-arm trials with the intention to account for unobserved confounders as well. Our purpose of considering this approach is to compare it to the novel approach of digital twins in terms of suitability for increasing the power while controlling the type I error.

The novel approach of digital twins is then treated in chapter 5, where we delve into the large theoretical background of using digital twins with ANCOVA models to estimate the average treatment effect. First, the concept of digital twins is presented, showing how these can be utilised in an ANCOVA model to estimate the average treatment effect. The goal of the chapter is to prove analytical results stating that the average treatment estimator incorporating digital twins in the analysis is the estimator that obtains the lowest asymptotic variance and therefore the largest asymptotic power among a large class of estimators. For that purpose, we begin by describing the concept of influence functions to prove several results regarding the asymptotic variance of various average treatment effect estimators utilising digital twins. We then use these asymptotic results to show that the efficiency gain of using digital twins can be exploited in order to decrease the required sample size of a trial.

In chapter 6, we conduct a simulation study in order to compare the performance of the SCA and digital twin approaches in terms of leveraging historical data. Specifically, we investigate how large efficiency gain can be achieved in different scenarios in a finite sample setting using different (machine learning) prognostic models to predict outcomes of digital twins. We also examine whether the type I error rate is controlled. We consider the performance of both methods in scenarios that are optimal in terms of the asymptotic results presented in chapter 5, and more realistic scenarios in terms of data distributions and model specifications. Additionally, we compare the methods to models not leveraging historical data as well as investigating sample size calculations based on approximation formulas described in chapter 3.

In chapter 7 we investigate whether digital twins can be used to increase power in an RCT intended for demonstrating efficacy of insulin products used for treating patients diagnosed with type 2 diabetes. Specifically, we use real world data originating from three RCTs previously conducted by Novo Nordisk A/S in order to examine the degree to which the number of patients required for recruitment to a current trial can be limited by using the digital twin approach.

Chapter 8 provides a discussion of the digital twins approach in regard to regulatory considerations, other approaches for leveraging historical data, and on some possible future developments

of the method. The chapter discusses a recent draft opinion from the European Medicines Agency (EMA) on the use of digital twins, another method used in a trial previously approved by the U.S. Food and Drug Administration (FDA), as well as our own reflections based on the analytical and empirical results contained in previous chapters of the thesis.

The appendix contains some of the most comprehensive parts of technical derivations necessary to reach conclusions throughout the thesis. In addition, further results from the simulation study and information regarding the data used in the case study are contained in the appendix. References to all specific parts of the appendix are made throughout the thesis.

Throughout the thesis, derivations will sometimes involve long equations, using several assumptions and properties throughout. Generally, all explanations are placed after the equations. All examples, remarks, lemmas, corollaries, propositions and theorems are finished by ◀, and proofs are finished by ■.

2 | Randomised Clinical Trials

In this chapter, we will describe how randomised clinical trials can be used to estimate the average treatment effect of some medical intervention. We will begin by describing the concept of randomised clinical trials, proceeding with introducing relevant notation and estimands describing the treatment effect, which are both used throughout the thesis. We will then continue with describing the benefits of randomisation in regard to establishing a causal relationship, and how different forms of randomisation can be conducted in practice. We will then briefly describe how ANOVA and ANCOVA models can be used to estimate the treatment effect in RCTs, and how using the ANCOVA model with adjustment for covariates can be beneficial.

Clinical trials are studies of the effect and safety of a medical intervention on a specific human population. Often clinical trials are used to assess the intervention compared to placebo or standard of care. Before the intervention is approved to proceed to human clinical trials, extensive preclinical trials in animals must be run. Among clinical trials, a *randomised clinical trial* (or *randomised controlled trial*) (RCT) is considered to generate the most reliable type of evidence. This is because RCTs are less biased than e.g. cohort studies due to its ability to distribute the confounders across the two groups, as we will show later.

An RCT proceeds by first assessing the eligibility of each patient according to inclusion and exclusion criteria specified in a pre-specified protocol. Thereafter the patients are randomly assigned to a treatment sequence, and throughout the study period, multiple research site visits are conducted to evaluate and record blood measurements, physical and cognitive condition etc. through a number of parameters. These visits are planned and registered in the protocol. Often in an RCT, both the patients, clinicians and data scientists are blinded to the treatment, meaning that group allocation is concealed [1, 2].

There are four phases of clinical trials, each with a specific purpose. Phase I studies are performed on a small sample of usually around 20-80 people and have a short course of treatment to evaluate the safety and dosage of a drug in the relevant population. Furthermore, it is used to assess the pharmacokinetics and pharmacodynamics of the drug which can be used to design the phase II and III studies. As an example, the half-life of a drug should be used to determine the duration of the study, to ensure that the participants reach a steady state. This is also used to determine the duration of the wash-out period necessary to avoid carryover effects between treatments in cross-over studies, where the patients are exposed to and switched between all treatment arms [1, 2]. Sometimes several phase I studies are performed for an intervention, and for some of these a control group might not be present. In these cases, the single-arm trials could potentially benefit from using historical data as an external control arm in order to get early indications of the drugs efficacy.

Phase II studies are longer studies usually including around 100-300 patients to further inves-

tigate the safety and tolerability, but also preliminary results about the therapeutic effect of the drug. It is also possible to investigate the short-term side effects of the drug. Phase I and II trials can also be used for early stage proof of concept to obtain preliminary evidence of efficacy for a clinically relevant endpoint [1, 3].

A phase III trial should instead provide pivotal information about the effectiveness and safety of a drug, and therefore the duration and sample size is increased. Around 300-3000 patients should usually participate, and the duration should be increased such that the long-term side effects can be assessed. Between these phases, an interim analysis is made to determine whether the drug should proceed to the next development state [1, 4]. After phase III, the drug is evaluated for approval by medical authorities as for example European Medicines Agency (EMA) or the U.S. Food and Drug Administration (FDA), and, after approval, the drug is moved to phase IV. The drug's effectiveness and safety is now monitored in a larger population for example by conducting a prospective cohort study.

2.1 Notation and Framework

In the setting of a two-arm clinical trial with n observations having a continuous endpoint (outcome), we let the stochastic variable Y_i denote the outcome of observation i (which we will refer to as a *participant*, *patient* or *subject*). We let Y_i^{pre} be the outcome measured before the intervention of interest (at *baseline*), and Y_i^{post} be the outcome measured after some prespecified time period receiving a medical intervention. Then by Y_i, we will usually mean the change from baseline of the endpoint, being the difference between Y^{post} and Y^{pre}. However, it can also denote the outcome after receiving the intervention, that is, $Y_i = Y_i^{\text{post}}$. Furthermore, we let X_i denote the stochastic $1 \times p$ row vector of baseline covariates, and denote by W_i the binary treatment assignment variable defined as $W_i := \mathbb{1}[\text{"patient } i \text{ is allocated to receive novel treatment"}]$.

Following this notation, observation i is a triple (X_i, W_i, Y_i), and the data set can be denoted as $(\mathbb{X}, \mathbb{W}, \mathbb{Y}) \in \mathcal{X}^n \times \{0,1\}^n \times \mathbb{R}^n$, where $\mathcal{X}$ is the sample space of the X_i's, allowing covariates to be continuous, binary and categorical. When doing linear analyses, we denote the design matrix by $\mathbb{D}$, specifying the relevant form in each case. Throughout the thesis, we will abuse notation and use f and p interchangeably to denote a probability density function, a probability mass function, or a mixture distribution function.

Unless otherwise specified, we assume that all data vectors (X_i, W_i, Y_i) are independent and identically distributed, meaning that the stochastic data vector of an arbitrary subject can be denoted as (X, W, Y). However, as we will see in section 2.2.1, for some cases of randomisation, correlation is introduced through the treatment assignment W_i. We say that patient i belongs to the *treatment group* if the realisation of W_i is $w_i = 1$, and that patient i belongs to the *control group* if $w_i = 0$. Thus, the treatment group consists of patients allocated to receive the medical intervention of interest, whereas the control group consists of patients allocated to receive placebo or standard of care. In the rest of this thesis, we will only consider two-arm and single-arm clinical trials.

In order to estimate the treatment effect, we will work in the framework of a Rubin causal model [5, 6], where we have two *potential outcomes*, $Y_i(1)$ and $Y_i(0)$, for patient i, denoting the stochastic variables describing the outcome of patient i under, respectively, treatment and no treatment. However, we will only observe the potential outcome corresponding to the treatment actually assigned to the patient. Formally, this means that we can denote any realisation of the outcome of patient i as $y_i = y_i(w_i)$. Since we observe n realisations of the random vector (X_i, W_i, Y_i), we will observe $n_1 := \sum_{i=1}^{n} w_i$ patients in the treatment group and $n_0 := \sum_{i=1}^{n} 1 - w_i$ patients in the control group. The treatment effect of patient i, which is unobserved, is then $y_i(1) - y_i(0)$. However, our goal is to estimate an average treatment effect across the n individuals.

We have three estimands to describe average treatment effects, namely the *average treatment effect* (ATE), the *average treatment effect of the treated* (ATT) and the *average treatment effect of the control group* (ATC). These are formally defined [5] as

$$\begin{aligned} \mathrm{ATE} &:= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ \mathrm{ATT} &:= \mathbb{E}\left[Y(1) \,\middle|\, W = 1\right] - \mathbb{E}\left[Y(0) \,\middle|\, W = 1\right] \\ \mathrm{ATC} &:= \mathbb{E}\left[Y(1) \,\middle|\, W = 0\right] - \mathbb{E}\left[Y(0) \,\middle|\, W = 0\right]. \end{aligned} \tag{2.1}$$

We can think of the ATE as the treatment effect across the whole sample consisting of n patients, that is, the average effect of moving the entire population from the control group to the treatment group. The ATT can be thought of as the average treatment effect of the patients that are treated. Similarly, ATC is the average treatment effect of the control group patients. Depending on the context, usually the ATE or the ATT is of interest [7].

Later, we will also use the conditional forms of these estimands, which are formally defined [5] as

$$\begin{aligned} \mathrm{CATE}(X) &:= \mathbb{E}[Y(1) \,|\, X] - \mathbb{E}[Y(0) \,|\, X] \\ \mathrm{CATT}(X) &:= \mathbb{E}\left[Y(1) \,\middle|\, X, W = 1\right] - \mathbb{E}\left[Y(0) \,\middle|\, X, W = 1\right] \\ \mathrm{CATC}(X) &:= \mathbb{E}\left[Y(1) \,\middle|\, X, W = 0\right] - \mathbb{E}\left[Y(0) \,\middle|\, X, W = 0\right], \end{aligned} \tag{2.2}$$

which are in general stochastic, being functions of the stochastic vector X.

In clinical trials, multiple outcomes are often of interest. These outcomes can e.g. be categorised as the primary endpoint of interest and secondary endpoints. A primary endpoint is a prespecified outcome that directly addresses the hypothesis of the trial, and it is confirmatory in the sense that it should be tested for significance. Furthermore, multiple secondary outcomes can be of interest. These can be confirmatory or exploratory, the latter in the sense that tests for statistical significance are not conducted.

The next sections describe how randomisation together with AN(C)OVA models can be used to estimate the average treatment effect measured by some continuous outcome.

2.2 Randomisation

This section is based on chapter 1 in [8].

The main goal of an efficacy trial is to collect data that enable describing a causal relationship between the intervention and the outcome to determine the ATE. Human response to an intervention W depends on patient characteristics X, and therefore the control group must be comparable to the treatment group, because differences between patient characteristics may otherwise confound the result. In regard to ensuring such comparability, the RCT design is considered to be the golden standard in clinical studies. Randomisation is the process of assigning participants to a treatment arm, such that there is equal probability for each person to be assigned to any given group.

However, randomisation sometimes poses and ethical dilemma, since the patients are assigned to either the treatment- or control group by chance and not by e.g. a physical judgement of their well being by a physician. This means that some patients will be exposed to a potentially beneficial intervention, but the intervention could also prove to be harmful or toxic. On the other hand, in trials involving treatments for a disease with no known treatment or cure, the control group could potentially be denied a beneficial treatment or cure. One may argue that since the trial often involves a novel drug, a physician would be in a state of equipoise, and therefore randomisation is not unethical. Contrary, a physician would often have some a priori knowledge about the treatments effect on a disease, especially in the later phase III studies. This poses another dilemma between what is best for the individual and what is best for the public health [8, pp. 9–12].

Having these considerations in mind, it may be argued that the use of historical cohorts of patients with no treatment or standard of care should be used as control groups, and that this would be more ethical than randomisation. As already mentioned, in some areas such as pediatric clinical trials and trials involving rare diseases, a more practical issue persists; within these areas, recruitment of large patient groups for clinical trials remains a major challenge [9, 10]. However, the direct use of historical controls as an external comparator can lead to skewed results, since the control group might differ from the treatment group in characteristics that may confound the study outcome. Both from a practical and ethical point of view, it can thus be beneficial to use other methods for incorporating historical data in order to decrease the required sample size in clinical trials [8, pp. 9–12]. It is possible to combine the benefits of randomisation with the use of historical controls, enabling the use of a smaller control group. This can ensure being able to more efficiently show efficacy of drugs and thus quicker let new drugs reach patients. Two ways of doing so will be presented in chapters 4 and 5.

In order to formally characterize the overall benefit of randomisation, we will define a *confounding covariate* X as a stochastic variable both influencing the exposure W and the outcome Y. This can result in estimating a false association between W and Y. Using Judea Pearl's causal calculus, a formal definition of confounding effects can be stated. In order to do so, we first we define the do-operator, which indicates an action or intervention. In a directed graphical model,

$\mathrm{do}(w)$ implies that we remove edges going into the point associated to W, but preserves the edges going out of this point. We then define *Bayesian conditioning* as $p(y \mid w)$ where w is an observed variable, and *causal conditioning* as $p\big(y \mid \mathrm{do}(w)\big)$, where the specific value w is forced. Then W and Y are not confounded if and only if

$$p\big(y \mid \mathrm{do}(w)\big) = p(y \mid w) \tag{2.3}$$

for all values of w and y [11, 12]. In the framework of causal calculus, randomisation thus entails that no edges are going into the point associated to W, so that Y and W are not confounded according to the definition (2.3).

The main reason for random assignment to a treatment sequence is to prevent selection bias, which occurs when the randomisation process is not conducted properly, e.g. when the allocation W is affected by covariates X, or when an e.g. deterministic allocation scheme prevents the allocation from being blinded for the analysts and medical professionals conducting the study [8, p. 79]. Randomisation does this by controlling the probability of the groups differing in regard to the observed and unobserved confounders, and, as we will show later, this probability tends to 0 when $n \to \infty$. In observational studies, this comparability between groups can be attempted for by matching observations based only on known covariates, which we will discuss further in chapter 4.

Formally, the randomisation in an RCT ensures that

$$\big(Y(0), Y(1)\big) \perp\!\!\!\perp W, \tag{2.4}$$

since the treatment allocation is assigned randomly. Thus, in this setting, we have

$$\begin{aligned} \mathbb{E}\big[Y(1)\big] - \mathbb{E}\big[Y(0)\big] &= \mathbb{E}\big[Y(1) \mid W = 1\big] - \mathbb{E}\big[Y(0) \mid W = 1\big] \\ &= \mathbb{E}\big[Y(1) \mid W = 0\big] - \mathbb{E}\big[Y(0) \mid W = 0\big] \end{aligned} \tag{2.5}$$

meaning that the estimands ATE, ATT and ATC in equation (2.1) are equal [6, 7].

In the following sections, we will describe how randomisation can be achieved in practice.

2.2.1 Complete and Forced Balance Randomisation

This section is based on chapters 3 and 4 in [8].

We denote by $\mathcal{W}_k = \{W_1, W_2, \ldots, W_k\}$ information on the treatment assignment of the first $k \leqslant n$ patients. If $W_1, W_2, \ldots, W_n$ are independent Bernoulli distributed variables with $\mathbb{P}(W_i = 1) = 1/2$ for each i, then the randomisation is called *complete randomisation*. This type of randomisation does not guarantee equally sized groups but there is no selection bias, since each patient is equally likely to be assigned to the treatment or control group, independent of what the previous patients have been assigned to. We then have that $n_1(n) = \sum_{i=1}^{n} W_i$ has a binomial distribution with mean $n/2$ and variance $n/4$. Then since $n_0(n) = n - n_1(n)$, it follows that

$D_n := n_1(n) - n_0(n) = 2n_1(n) - n$ has mean 0 and variance n. This implies that the mean difference between the group sizes is 0.

Using the central limit theorem, $n_1(n)$ is asymptotically normally distributed, which thus also holds for D_n, so that for large n, $\frac{1}{\sqrt{n}}D_n \approx \mathcal{N}(0,1)$. We can use this to determine a formula that describes the approximate probability of imbalances of size $\varepsilon > 0$. Specifically, we have for large n, that

$$\begin{aligned}\mathbb{P}\left(|D_n| > \varepsilon\right) &= \mathbb{P}\left(\frac{1}{\sqrt{n}}D_n > \frac{\varepsilon}{\sqrt{n}}\right) + \mathbb{P}\left(\frac{1}{\sqrt{n}}D_n < \frac{-\varepsilon}{\sqrt{n}}\right) \approx 1 - \Phi\left(\frac{\varepsilon}{\sqrt{n}}\right) + \left(\Phi\left(\frac{-\varepsilon}{\sqrt{n}}\right)\right) \\ &= 1 - \Phi\left(\frac{\varepsilon}{\sqrt{n}}\right) + \left(1 - \Phi\left(\frac{\varepsilon}{\sqrt{n}}\right)\right) = 2\left(1 - \Phi\left(\frac{\varepsilon}{\sqrt{n}}\right)\right), \end{aligned} \tag{2.6}$$

for Φ denoting the standard normal cumulative distribution function. The formula can be used in designing trials with a complete randomisation scheme. From the formula, we see that the probability of an absolute group size imbalance D_n of any fixed size ε converges to 1. However, similar calculations show that the probability of a relative imbalance D_n/n of arbitrary fixed size $\varepsilon > 0$ can be expressed as $2\left(1 - \Phi\left(\varepsilon\sqrt{n}\right)\right)$, which converges to 0.

Even though the complete randomisation scheme does not necessarily create equally sized groups, the treatment estimate in a linear model remains unbiased, but imbalance decreases the power as we will see later. To insure balanced in group sizes, a *forced balance randomisation* can be used, where the number of patients assigned to treatment 1 and 0 is exactly $n/2$. Due to the eligibility criteria checked before randomisation, the total number of patients randomised is often not known at the stage of trial design, and therefore these randomisation schemes can be used in blocks, where patients are randomised within blocks. Several forced balance randomisation schemes exist, and an example of such a randomisation scheme is the *random allocation scheme*.

Example 2.2.1. Random allocation scheme
For the random allocation scheme, an allocation rule is defined as

$$\mathbb{E}\left[W_j \,|\, \mathcal{W}_{j-1}\right] = \frac{\frac{n}{2} - n_1(j-1)}{n - (j-1)}, \quad j = 2, 3, \ldots, n, \tag{2.7}$$

with $\mathbb{E}[W_1] = \mathbb{P}(W_1 = 1) = 1/2$. We note that the randomisation rule is itself a random variable with expected value

$$\mathbb{P}(W_j = 1) = \mathbb{E}[W_j] = \mathbb{E}\left[\mathbb{E}\left[W_j \,|\, \mathcal{W}_{j-1}\right]\right] = \mathbb{E}\left[\frac{\frac{n}{2} - \sum_{i=1}^{j-1} W_i}{n - (j-1)}\right] = \frac{\frac{n}{2} - (j-1)\frac{1}{2}}{n - (j-1)} = \frac{1}{2}, \tag{2.8}$$

where the fourth equality follows by iteratively solving the equation for $j = 2, 3, \ldots, n$. Having e.g. $n = 100$, and the first 49 patients being allocated as $n_1(49) = 28$ and $n_0(49) = 21$,

patient number 50 will be allocated to treatment 1 with probability $(50-28)/51 = 22/51$. This randomisation rule have some limitations. For example when $n/2$ patients have been allocated to a treatment, the rest of the patients will have allocation rule all equal to 1 or 0 meaning that the conditional allocations are deterministic. Thereby a selection bias can occur. ◀

Forced balance randomisation will in general control the probability of the covariates differing by a specific amount. We start by deriving this for the case of one-dimensional covariate vectors. Let the covariates $X_1, X_2, \ldots, X_n$ for each patient be independent of the treatment assignment and mutually independent with mean μ and variance σ^2. Furthermore, we let $\mathbb{P}(W = 1) = \mathbb{E}[W] = \pi_1 \in]0,1[$ and $\overline{X}_1 = \sum_{i=1}^n \frac{W_i X_i}{n\pi_1}$ and $\overline{X}_0 = \sum_{i=1}^n \frac{(1-W_i)X_i}{n(1-\pi_1)}$. Under forced balance randomisation, $n\pi_1 = n_1$ and $n(1-\pi_1) = n_0$, so $\overline{X}_1$ and $\overline{X}_0$ are sample means of the covariates in the treatment and control group, respectively. Using linearity of the (conditional) expected value, we get

$$\begin{aligned}\mathbb{E}\left[\overline{X}_1 - \overline{X}_0\right] &= \mathbb{E}_{\mathcal{W}_n}\left[\mathbb{E}\left[\sum_{i=1}^n \frac{W_i X_i}{n\pi_1} - \sum_{i=1}^n \frac{(1-W_i)X_i}{n(1-\pi_1)} \,\middle|\, W_1, W_2, \ldots, W_n\right]\right] \\ &= \sum_{i=1}^n \frac{\mathbb{E}_{\mathcal{W}_n}[W_i]\mu}{n\pi_1} - \sum_{i=1}^n \frac{\mathbb{E}_{\mathcal{W}_n}[1-W_i]\mu}{n(1-\pi_1)} = \sum_{i=1}^n \frac{\mu}{n} - \sum_{i=1}^n \frac{\mu}{n} = 0.\end{aligned} \tag{2.9}$$

Therefore the expected value of the differences in covariate values is 0, no matter the value of π_1. From the above derivation it is seen that this is also true when relaxing the assumption of $\mu_i = \mu$ for all i, with μ_i denoting the expected value of the covariate belonging to the ith subject.

In case of deterministic allocation, which we present in the next example, the expected difference in group means is also 0 when $\mu_i = \mu$ for all i. The assumption that $\mu_i = \mu$ for all i is sometimes unrealistic in practice since e.g. time trends could occur. When the assumption is violated, deterministic allocation does not ensure equal expected group means. This is due to the fact that this allocation scheme does not satisfy the general assumptions worded above, as π_1 needs to be defined differently as a consequence of the determinism of the treatment assignment.

Example 2.2.2. Deterministic allocation scheme
From equation (2.9) we see that if the W_i's are deterministic, and we in this case define $\pi_1 = \frac{1}{n}\sum_{i=1}^n W_i \in]0,1[$ as the deterministic allocation ratio, then for example given that the first $n\pi_1$ patients are allocated to treatment 1 and the last $n(1-\pi_1)$ patients to treatment 0, we have

$$\begin{aligned}\mathbb{E}\left[\overline{X}_1 - \overline{X}_0\right] = \mathbb{E}\left[\sum_{i=1}^n \frac{W_i X_i}{n\pi_1} - \sum_{i=1}^n \frac{(1-W_i)X_i}{n(1-\pi_1)}\right] &= \sum_{i=1}^n \frac{W_i\mu}{n\pi_1} - \sum_{i=1}^n \frac{(1-W_i)\mu}{n(1-\pi_1)} \\ &= \sum_{i=1}^{n\pi_1} \frac{\mu}{n\pi_1} - \sum_{i=1}^{n(1-\pi_1)} \frac{\mu}{n(1-\pi_1)} = 0.\end{aligned} \tag{2.10}$$

Thus, we have balanced groups in this case, but in practice such a deterministic allocation scheme is not used since this will disturb the blinding. Moreover, if $\mu_i \neq \mu$, the expected difference would not equal 0. ◀

Returning to the general case of forced balance randomisation, we have by the law of total variance that

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\overline{X}_1 - \overline{X}_0\right] = \mathbb{E}_{\mathcal{W}_n}\left[\mathbb{V}\mathrm{ar}\left[\sum_{i=1}^n \frac{W_i X_i}{n\pi_1} - \sum_{i=1}^n \frac{(1-W_i)X_i}{n(1-\pi_1)} \,\middle|\, W_1, W_2, \ldots, W_n\right]\right] \\ + \mathbb{V}\mathrm{ar}_{\mathcal{W}_n}\left[\mathbb{E}\left[\sum_{i=1}^n \frac{W_i X_i}{n\pi_1} - \sum_{i=1}^n \frac{(1-W_i)X_i}{n(1-\pi_1)} \,\middle|\, W_1, W_2, \ldots, W_n\right]\right].\end{aligned} \tag{2.11}$$

We begin by regarding the last term, which is equal to

$$\begin{aligned}&\mathbb{V}\mathrm{ar}_{\mathcal{W}_n}\left[\mathbb{E}\left[\sum_{i=1}^n \frac{W_i X_i}{n\pi_1} - \sum_{i=1}^n \frac{(1-W_i)X_i}{n(1-\pi_1)} \,\middle|\, W_1, W_2, \ldots, W_n\right]\right] \\ &= \mathbb{V}\mathrm{ar}_{\mathcal{W}_n}\left[\sum_{i=1}^n \frac{W_i \mu}{n\pi_1} - \sum_{i=1}^n \frac{(1-W_i)\mu}{n(1-\pi_1)}\right] = \mathbb{V}\mathrm{ar}_{\mathcal{W}_n}\left[\mu\sum_{i=1}^n \frac{W_i}{n\pi_1} - \mu\sum_{i=1}^n \frac{(1-W_i)}{n(1-\pi_1)}\right] \\ &= \mathbb{V}\mathrm{ar}_{\mathcal{W}_n}\left[\mu - \mu\right] = 0,\end{aligned} \tag{2.12}$$

where we have used linearity of the conditional expected value in the first equality while regarding each W_i as constant and in the last equality we use that forced balance randomisation is assumed, such that the sums each equal 1. We can then determine the variance of the group difference as

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\overline{X}_1 - \overline{X}_0\right] &= \mathbb{E}_{\mathcal{W}_n}\left[\mathbb{V}\mathrm{ar}\left[\sum_{i=1}^n \frac{W_i X_i}{n\pi_1} - \sum_{i=1}^n \frac{(1-W_i)X_i}{n(1-\pi_1)} \,\middle|\, W_1, W_2, \ldots, W_n\right]\right] \\ &= \frac{\sigma^2}{(n\pi_1)^2}\sum_{i=1}^n \mathbb{E}_{\mathcal{W}_n}\left[W_i^2\right] + \frac{\sigma^2}{\left(n(1-\pi_1)\right)^2}\sum_{i=1}^n \mathbb{E}_{\mathcal{W}_n}\left[(1-W_i)^2\right] \\ &= \frac{\sigma^2}{(n\pi_1)^2}\sum_{i=1}^n \mathbb{E}_{\mathcal{W}_n}\left[W_i\right] + \frac{\sigma^2}{\left(n(1-\pi_1)\right)^2}\sum_{i=1}^n \mathbb{E}_{\mathcal{W}_n}\left[(1-W_i)\right] \\ &= \frac{\sigma^2}{n\pi_1} + \frac{\sigma^2}{n(1-\pi_1)} = \frac{\sigma^2}{n\pi_1(1-\pi_1)}.\end{aligned} \tag{2.13}$$

In the second equality we use that by conditioning on $\mathcal{W}_n$ we can regard each W_i as a constant and thereby use linearity of variance since the X_i's are independent across $i = 1, 2, \ldots, n$. Taking the variance on each term we obtain W_i^2 times the variance, but since W_i is either 1 or 0, we have

$W_i^2 = W_i$ and $(1-W_i)^2 = 1-W_i$, which is used in the third equality. Then by (2.9), (2.13) and Chebyshev's inequality, we have that for any $\varepsilon > 0$

$$\mathbb{P}\left(\left|\overline{X}_1 - \overline{X}_0\right| \geqslant \varepsilon\right) \leqslant \frac{1}{\varepsilon^2}\left(\frac{\sigma^2}{n\pi_1(1-\pi_1)}\right), \tag{2.14}$$

which goes to 0 as $n \to \infty$. Therefore, for large enough n, the probability of the difference in covariate values being larger than ε becomes negligible. We also note that for any n, the obtained bound on the difference is smallest when $\pi_1 = 1/2$.

Remark. Rosenberger and Lachin [8] state in the beginning of their section 4.2, that the variance in equation (2.12) is equal to 0 for different means μ_i. However, after getting in touch with the authors, they confirmed that this does not seem to hold in general. ◀

The derivation was done for X_i being one dimensional. The same arguments can be used for X_i being of dimension p, and replacing σ^2 by a positive definite matrix Σ and μ by a vector of the mean values. The same arguments hold as for the one-dimensional case, with the same derivations until equation (2.13), in which case we get, analogously, that

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\overline{X}_1 - \overline{X}_0\right] &= \mathbb{E}_{\mathcal{W}_n}\left[\mathbb{V}\mathrm{ar}\left[\sum_{i=1}^n \frac{W_i X_i}{n\pi_1} - \sum_{i=1}^n \frac{(1-W_i)X_i}{n(1-\pi_1)} \,\middle|\, W_1, W_2, \ldots, W_n\right]\right] \\
&= \frac{\Sigma}{(n\pi_1)^2}\sum_{i=1}^n \mathbb{E}_{\mathcal{W}_n}\left[W_i^2\right] + \frac{\Sigma}{\left(n(1-\pi_1)\right)^2}\sum_{i=1}^n \mathbb{E}_{\mathcal{W}_n}\left[(1-W_i)^2\right] \\
&= \cdots = \frac{\Sigma}{n\pi_1(1-\pi_1)}.
\end{aligned} \tag{2.15}$$

From the multivariate Chebyshov's inequality, we obtain that for any $\varepsilon > 0$

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{\left(\overline{X}_1 - \overline{X}_0\right)^\top \Sigma^{-1}\left(n\pi_1(1-\pi_1)\right)\left(\overline{X}_1 - \overline{X}_0\right)} \geqslant \varepsilon\sqrt{n\pi_1(1-\pi_1)}\right) \\
= \mathbb{P}&\left(\sqrt{\left(\overline{X}_1 - \overline{X}_0\right)^\top \Sigma^{-1}\left(\overline{X}_1 - \overline{X}_0\right)} \geqslant \varepsilon\right) \leqslant \frac{1}{\varepsilon^2}\frac{p}{n\pi_1(1-\pi_1)}.
\end{aligned} \tag{2.16}$$

As for the one-dimensional case, we obtain that the upper bound goes to 0 as $n \to \infty$, and in practice for large n with $p \ll n$, the probability becomes negligible.

2.2.2 Stratified Randomisation

This section is based on chapter 7 in [8].

In clinical trials we observe covariates, which are presumed to be associated with the outcome for each patient. Even though randomisation seeks to minimize the heterogeneity of the groups in order to remove bias, imbalances can occur in practice. For example in multicenter trials, the difference in clinical center could cause heterogeneity of the groups. This is due to the clinics differing in demographics, patient populations and compliance in regard to the protocol. Therefore the number of patients randomised to each group within a clinic should be balanced to minimise bias. In the event that there is imbalance in regard to a prognostic categorical covariate, a *stratified-adjusted analysis* can be used by adjusting for this covariate. We note that here that the term stratified does not refer to estimating different treatment effects for different levels of the categorical covariate, which could be obtained by a *subgroup analysis* within strata, or including an interaction term in the ANCOVA model introduced later in equation (2.20). The trial design can also ensure balance in regard to that specific covariate through *stratified randomisation*.

Stratified randomisation is a part of the trial design and takes place before the actual randomisation of subjects. The participants are grouped according to their values of some prognostic covariates and then randomised within these strata. For instance, when stratifying in regard to clinic (total of 10 clinics) and sex, there should be a randomisation sequence (obtained from an allocation rule) for each stratum and therefore a total of 20 randomisation sequences. When a forced balance randomisation is used, the probability of each subject being assigned to treatment 1 or 0 only depends on the prior assignments of subjects in the same stratum. For example, a female at clinic 3 would be randomised only depending on the assignment of other females at clinic 3. For some randomisation schemes we only have asymptotic assurance of equally sized treatment- and control groups, and therefore there is a positive probability of having imbalances in group sizes. This means that when we randomise in each stratum, these imbalances become additive and this can result in larger overall imbalance. For few large strata, the probability of this occurring is smaller. When using the random allocation rule, there will be no imbalance in the group sizes of each stratum unless the number of subjects in a stratum is not even.

A stratified-adjusted analysis is done at the end of the study when all the data is collected. This can be carried out in combination with or without a stratified randomisation. Here the prognostic covariates are incorporated in the analysis. The procedure is determined by the specific model used, and we will discuss this for the ANCOVA model described in section 2.3. Some argue that it suffices to conduct a stratified-adjusted analysis without stratified randomisation, since the randomisation itself seeks to avoid imbalances in the covariate values, at least when n is large, and since small imbalances do not affect the results by much when they are adjusted for in the analysis [8, p. 135].

2.3 AN(C)OVA in Randomised Clinical Trials

Analysis of variance (ANOVA) is a classical method of estimating the average treatment effect in RCTs across the two treatment groups. In general, ANOVA is a collection of statistical models used to analyze the difference between means of groups. In the context of RCTs, one-way ANOVA can be used to estimate the treatment effect as the coefficient β_W in the linear model

$$Y_i = \beta_0 + W_i\beta_W + \varepsilon_i \tag{2.17}$$

where the W_i's are independent of $\varepsilon_i \sim \mathcal{N}\left(0, \sigma_Y^2\right)$ and the ε_i's are mutually independent. That is, we assume that the model

$$Y \mid W \sim \mathcal{N}\left(\beta_0 + W\beta_W, \sigma_Y^2\right) \tag{2.18}$$

is appropriate. The maximum likelihood estimate (MLE) $\widehat{\beta}_0$ of the intercept is the mean outcome value of subjects in the control group, and the mean outcome in the treatment group is $\widehat{\beta}_0 + \widehat{\beta}_W$ such that the ML estimate $\widehat{\beta}_W$ is the difference of means in the two study arms, as seen by appendix A.1.

From the ANOVA model we can estimate ATE by $\widehat{\beta}_W$. This can be seen since for an RCT

$$\begin{aligned} \text{ATE} &= \mathbb{E}\left[Y(1)\right] - \mathbb{E}\left[Y(0)\right] \\ &= \mathbb{E}\left[Y(1) \mid W = 1\right] - \mathbb{E}\left[Y(0) \mid W = 0\right] \\ &= \mathbb{E}\left[Y \mid W = 1\right] - \mathbb{E}\left[Y \mid W = 0\right] = (\beta_0 + \beta_W) - \beta_0 = \beta_W, \end{aligned} \tag{2.19}$$

where the second equality follows from the property in equation (2.4) [6, 7], the third equality follows from the definition of $Y(W)$ and the fourth equality follows from taking conditional expectations in the model specification (2.18). From equation (2.5), we get that $\widehat{\beta}_W$ can be interpreted as ATE, ATT and ATC.

Analysis of covariance (ANCOVA) can be utilised as a method of estimating the average treatment effect similarly to the ANOVA using a normal linear model. However, the ANCOVA model does so while adjusting for prognostic baseline covariates, that is, baseline covariates that are anticipated to be associated with the primary endpoint. For ease of notation, we will assume that the p-dimensional vector of baseline covariates X_i for subject i consists of all prognostic baseline covariates. Formally we seek to estimate β_W in the linear model

$$Y_i = \beta_0 + W_i\beta_W + X_i\beta_X + \varepsilon_i, \tag{2.20}$$

where $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$ are mutually independent, and we assume that W_i and X_i are mutually independent (due to randomisation) and each independent of the ε_i's. That is, we assume that the model

$$Y \mid W, X \sim \mathcal{N}\left(\beta_0 + W\beta_W + X\beta_X, \sigma^2\right) \tag{2.21}$$

is appropriate. Furthermore, not including the treatment assignment variable in the design matrix of the linear model, we assume that the columns of the $n \times (p+1)$ design matrix $[1 \ \mathbb{X}]$ has rank $p+1$. We notice that from this model specification, we assume a homogeneous treatment effect across covariate values since we have no interaction effects between treatment and other covariates. From a geometric point of view, we obtain parallel hyper-planes across the treatment categories, since β_X is the same across the treatment groups. From this model specification, taking conditional expectation in case $W = 0$ and $W = 1$, we get $\beta_W = \mathbb{E}\left[Y(1) \mid X\right] - \mathbb{E}\left[Y(0) \mid X\right] = \text{CATE}$, which in this case is not stochastic since we do not model interaction effects, whereby the terms including X cancel out. Taking unconditional expectation we also obtain $\beta_W = \text{ATE}$. Later, in section 3.5 we show that in an RCT, even though assumptions of the ANCOVA model are violated, the ANCOVA ATE estimator is consistent.

In appendix A.3 we derive an explicit expression of the ANCOVA ML estimator of an equivalent, reparametrised model. Specifically, we see that the ANCOVA ML estimator in equation (A.15), like the ANOVA ML estimator, contains the difference of group means between the two study arms. However in this case, we adjust for the difference of group means for all covariates and weigh these by their parameters, respectively.

The adjustment serves as a method to account for chance imbalances in the baseline covariates across the randomised study arms. Adjusting for highly prognostic covariates has furthermore been shown to lead to substantial increase in power over the ANOVA method in simulation studies based on previously conducted RCTs [13]. In chapter 3, we will delve into the theoretical background of how this gain in power is achieved. Specifically, we will show that the power increases when the standard error decreases. The improvement in power thus occurs since more of the variation in the outcome between patients is explained by the added covariates, leading to a smaller residual standard error, and thereby a smaller standard error of the treatment estimate, as we will show in section 2.3.1. Therefore, even when there, by chance, is a perfect or near perfect balance in the baseline covariates across the study arms, adjusting for covariates using ANCOVA can increase the power.

Some practical considerations need to be taken when estimating the ATE in a clinical trial. As an example, even though it is the intention that all patients allocated to the treatment group actually receives the treatment of interest, in practice a number of participants will withdraw from the study or deviate from the protocol, e.g. due to adverse events. An analysis method relevant when deviations or withdrawals are present is the *intention-to-treat* principle, where all randomised patients are included in the analysis no matter if any intercurrent event or deviations from the protocol happens. An intercurrent event is an event preventing the patient from being on the assigned intervention. In the analysis, we will then consider these patients as belonging to the treatment- and control groups according to which group the participant was originally assigned to. If a patient withdraws from the study, investigators should still try to collect the measurements indicated by the protocol. The intention-to-treat principle is different from the *per protocol* and *as-treated* principles, where patients are considered as belonging to the group according to the actual treatment received. According to the per protocol principle, patients deviating from the protocol should be excluded from the analysis, as opposed to the as-treated principle.

For the two latter procedures, the benefits of randomisation are not preserved, meaning that we do not ensure expected balance between the groups in regard to prognostic covariates. Thus, when patients are not analysed according to their intended treatment assignment, the risk of bias will increase when estimating the ATE. Specifically, these approaches can lead to spuriously concluding that the intervention is effective [14]. In a clinical setting, this would be particularly alarming since approval could lead to patients being exposed to a potentially ineffective drug.

Conversely, the intention-to-treat principle does preserve the benefits of randomisation. There can still be substantial non-adherence to the protocol, but in this case, the intention-to-treat principle will in worst case underestimate the magnitude of the effect of the intervention relative to the situation of perfect adherence. This is due to empirical evidence suggesting that, even when controlling for known prognostic factors, patients tend to do better when they adhere to the protocol regardless of whether they receive treatment or not. Thus, if an intervention is truly effective, the intention-to-treat principle will still not allow for inflated type I errors. In this way, the principle provides an unbiased estimate of the ATE at the level of adherence in the study [14], making the results comparable to the effect of a drug in a real-world setting, where patients are not perfectly compliant.

2.3.1 Benefits of Covariate Adjustment in RCTs

Randomisation ensures independence between the treatment allocation and other (observed and unobserved) covariates. This means that when we wish to find the treatment effect, we can choose to control for other covariates than treatment without running the risk of introducing bias to the treatment effect estimate. In fact, controlling for other covariates through ANCOVA is quite common in clinical studies. This practice is also known as covariate adjustment. The hope is to reduce the variance of the β_W estimate. For a normal linear model with design matrix $\mathbb{D} = [1\ \mathbb{W}\ \mathbb{X}]$, we have

$$\mathbb{V}\text{ar}\left(\widehat{\beta}\right) = \sigma^2\left(\mathbb{D}^\top\mathbb{D}\right)^{-1} \tag{2.22}$$

for the MLE of $\beta = (\beta_0, \beta_W, \beta_X)^\top$. When introducing (more) covariates in $\mathbb{X}$, we would reduce the residual error σ^2, thus reducing the variance as desired. However, the matrix factor in equation (2.22) also changes when introducing covariates, thus not guaranteeing a reduction in variance. However, as we will now show, we expect a reduction in variance of $\widehat{\beta}_W$ when introducing (more) covariates in $\mathbb{X}$, when we are in the setup of an RCT. However, we run the risk of overfitting and thereby underestimating the true variance σ^2.

Specifically, we start by showing that under an RCT, we have expected orthogonality of the demeaned W and the demeaned covariates. Under exact orthogonality, we can show then that the ANOVA treatment estimator is unbiased, and that we expect the same estimator from the ANCOVA model, but with lower variance.

We can demean W and the covariates and still obtain equivalent results. This can be seen by considering a model which is transformed with the orthogonal projection of the 1-column of

the design matrix. From the Frisch-Waugh-Lovell (FWL) theorem [15, p. 69], we get for the orthogonal projection $M_0 = I_n - P_0 = I_n - 1(1^\top 1)^{-1}1^\top$ of the 1-vector, that

$$M_0\mathbb{Y} = M_0\mathbb{W}\beta_W + M_0\mathbb{X}\beta_X + \varepsilon \tag{2.23}$$

provides a regression with the same ML estimate and errors as the model in equation (2.20). Hence, properties of this model also hold for the non-transformed model. The model is centralised in the sense that for any conformable vector v, $M_0 v := \widetilde{v} = v - \overline{v}$, where $\overline{v}$ denotes the empirical mean of the entries in v.

In practice, we do not have complete orthogonality between $\widetilde{\mathbb{W}}$ and $\widetilde{\mathbb{X}}$, but we expect the orthogonality to be more precise when a sufficient number of patients are included and hence randomised. Using that $\mathbb{E}\left[\widetilde{X}\right] = 0 = \mathbb{E}\left[\widetilde{W}\right]$, we have that

$$\mathbb{C}\mathrm{ov}\left(\widetilde{W}, \widetilde{X}\right) = \mathbb{E}\left[\widetilde{W}\widetilde{X}\right] + \mathbb{E}\left[\widetilde{W}\right]\mathbb{E}\left[\widetilde{X}\right] = \mathbb{E}\left[\widetilde{W}\widetilde{X}\right]. \tag{2.24}$$

In the case of an RCT, the treatment allocation is randomised, so we have independence between X and W, yielding $\mathbb{C}\mathrm{ov}(W, X) = 0$, which means by equation (2.24), that $\mathbb{E}\left[\widetilde{W}\widetilde{X}\right] = 0$. That is, in case of an RCT, we have expected orthogonality between $\widetilde{\mathbb{W}}$ and $\widetilde{\mathbb{X}}$.

The following example [15, pp. 112-114] illustrates that, in general for linear models, including too many covariates never causes bias if all the regressors, from the linear data generating process, are present. Only the efficiency of the estimator is affected, unless overspecified regressors are orthogonal to the true data generating regressors.

Example 2.3.1. Overspecification in linear models
Suppose some stochastic outcome $\mathbb{Y}$ is generated by a stochastic design matrix $\mathbb{X}_1$ containing p_1 covariates, by the linear relation

$$\mathbb{Y} = \mathbb{X}_1\beta_1 + \varepsilon, \tag{2.25}$$

where the covariates are independent of $\varepsilon \sim \mathcal{N}_n\left(0, \sigma^2 I_n\right)$ and we denote the corresponding ML estimate of β_1 as $\widehat{\beta}_1^{\mathrm{TRUE}}$. We say that the model is *overspecified* if we try to model the data by additionally using some other design matrix $\mathbb{X}_2$ of dimension p_2 as

$$\mathbb{Y} = \mathbb{X}_1\beta_1 + \mathbb{X}_2\beta_2 + \varepsilon. \tag{2.26}$$

Since the overspecified model is a special case of the correct model with $\beta_2 = 0$, the correct model is contained within the overspecified model.

The FWL theorem gives that the ML estimate of β_1 in the regression (2.26) of Y on X_1 and X_2 is numerically identical to the ML estimate of β_1 in the model

$$M_2\mathbb{Y} = M_2\mathbb{X}_1\beta_1 + \varepsilon \tag{2.27}$$

with the same error terms, and where $M_2 = I_n - \mathbb{X}_2(\mathbb{X}_2^\top \mathbb{X}_2)\mathbb{X}_2^\top$ is the orthogonal projection of the column space of $\mathbb{X}_2$. Thus, estimating the model (2.27), we can express the ML estimate of β_1 as

$$\widehat{\beta}_1^{\text{OS}} = \left((M_2\mathbb{X}_1)^\top M_2\mathbb{X}_1\right)^{-1}(M_2\mathbb{X}_1)^\top M_2\mathbb{Y} = \left(\mathbb{X}_1^\top M_2\mathbb{X}_1\right)^{-1}\mathbb{X}_1^\top M_2\mathbb{Y}. \tag{2.28}$$

This estimator is unbiased since

$$\begin{aligned}\mathbb{E}\left[\widehat{\beta}_1^{\text{OS}}\right] &= \mathbb{E}\left[(\mathbb{X}_1^\top M_2\mathbb{X}_1)^{-1}\mathbb{X}_1^\top M_2(\mathbb{X}_1\beta_1 + \varepsilon)\right] = \beta_1 + \mathbb{E}\left[(\mathbb{X}_1^\top M_2\mathbb{X}_1)^{-1}\mathbb{X}_1^\top M_2\varepsilon\right] \\ &= \beta_1 + \mathbb{E}\left[(\mathbb{X}_1^\top M_2\mathbb{X}_1)^{-1}\mathbb{X}_1^\top M_2\,\mathbb{E}\left[\varepsilon \mid \mathbb{X}_1, \mathbb{X}_2\right]\right] = \beta_1\end{aligned} \tag{2.29}$$

using law of total expectation and the independence between covariates and the error terms. The variance of the estimator is given by

$$\mathbb{V}\text{ar}\left(\widehat{\beta}_1^{\text{OS}}\right) = \sigma^2(\mathbb{X}_1^\top M_2\mathbb{X}_1)^{-1}. \tag{2.30}$$

As we will now show, the estimators in the overspecified model are inefficient. Specifically, the variances of the regression coefficient estimators are smaller when estimating from the correctly specified model than when overspecifying the model with regressors that are not orthogonal to the true regressors. We can prove this by showing that $\mathbb{V}\text{ar}\left(\widehat{\beta}_1^{\text{OS}}\right) - \mathbb{V}\text{ar}\left(\widehat{\beta}_1^{\text{TRUE}}\right)$ is positive semi-definite, meaning that all diagonal elements of this difference are non-negative. This is equivalent to showing that $\mathbb{V}\text{ar}\left(\widehat{\beta}_1^{\text{TRUE}}\right)^{-1} - \mathbb{V}\text{ar}\left(\widehat{\beta}_1^{\text{OS}}\right)^{-1}$ is positive semi-definite, which follows from lemma A.5.1, which is stated and proved in appendix A.5. If we let $P_2 = I_n - M_2$ denote the orthogonal projection onto the column space of $\mathbb{X}_2$, we indeed see that

$$\begin{aligned}\mathbb{V}\text{ar}\left(\widehat{\beta}_1^{\text{TRUE}}\right)^{-1} - \mathbb{V}\text{ar}\left(\widehat{\beta}_1^{\text{OS}}\right)^{-1} &= \sigma^{-2}(\mathbb{X}_1^\top\mathbb{X}_1) - \sigma^{-2}(\mathbb{X}_1^\top M_2\mathbb{X}_1) = \sigma^{-2}\left(\mathbb{X}_1^\top(I_n - M_2)\mathbb{X}_1\right) \\ &= \sigma^{-2}(\mathbb{X}_1^\top P_2\mathbb{X}_1) = \sigma^{-2}(\mathbb{X}_1^\top P_2^\top P_2\mathbb{X}_1) \\ &= \sigma^{-2}(\mathbb{X}_1P_2)^\top(\mathbb{X}_1P_2)\end{aligned} \tag{2.31}$$

is positive semi-definite. Note here that the true residual variances σ^2 from the correctly specified model and the overspecified model are the same, since the true β_2 is 0. Moreover, in the case where the overspecified regressors are orthogonal to the true data generating regressors, we have $P_2\mathbb{X}_1 = 0$, making the overspecified estimator efficient. ◀

In the case of an RCT, using demeaned data, we have expected ortogonality, and therefore we expect that overspecification will not affect the effiency of the estimator by much. Specifically, using example 2.3.1 with $\mathbb{X}_1 = \mathbb{W}$, we have expected orthogonality with any covariates $\mathbb{X}_2$, overspecified or not.

In the opposite situation, that is, in the case of underspecification, where we fail to include some data generating regressors, we risk encountering biased and inconsistent estimators, as we will illustrate for the linear model in the next example.

Example 2.3.2. Underspecification in linear models
Consider the reverse situation than in example 2.3.1, that is, where (2.26) is the data generating process while we try to estimate it using the model in (2.25). We regard this model as misspecified, since the specified model does not collapse to the data generating model for any choice of parameters. Then we see that our ML estimate is given by

$$\begin{aligned}\widehat{\beta}_1^{\text{US}} &= (\mathbb{X}_1^\top \mathbb{X}_1)^{-1}\mathbb{X}_1^\top \mathbb{Y} = (\mathbb{X}_1^\top \mathbb{X}_1)^{-1}\mathbb{X}_1^\top (\mathbb{X}_1\beta_1 + \mathbb{X}_2\beta_2 + \varepsilon) \\ &= \beta_1 + (\mathbb{X}_1^\top \mathbb{X}_1)^{-1}\mathbb{X}_1^\top \mathbb{X}_2\beta_2 + (\mathbb{X}_1^\top \mathbb{X}_1)^{-1}\mathbb{X}_1^\top \varepsilon.\end{aligned} \tag{2.32}$$

In expectation this is equal to $\beta_1 + \mathbb{E}\left[(\mathbb{X}_1^\top \mathbb{X}_1)^{-1}\mathbb{X}_1^\top \mathbb{X}_2\beta_2\right]$ using law of total expectation and the assumption of independence between regressors and the error term on the third term. Thus, in the case of underspecification, the β_1 estimator is biased unless the omitted regressors are orthogonal to the ones used in the model, that is, if $\mathbb{X}_1^\top \mathbb{X}_2 = 0$. We refer to this property as *omitted variable bias*. In the case of non-orthogonality, the second term is non-zero even asymptotically, so in addition to the estimator being biased, it is inconsistent as well [15, pp. 114-115]. ◀

Recall that in the setting of an RCT, using demeaned data, we have expected orthogonality between $\widetilde{\mathbb{X}}$ and $\widetilde{\mathbb{W}}$. From example 2.3.2 we then see that under orthogonality, if data is generated by a linear model from a number of covariates, the ATE estimates obtained from an ANOVA model or an underspecified ANCOVA model is unbiased, even though these are underspecified. In the following example, we see that under orthogonality, the ANCOVA and ANOVA estimators coincide.

Example 2.3.3. Relation between ANOVA and ANCOVA ATE estimators
Assuming exact orthogonality $\mathbb{X}^\top \mathbb{W} = 0$, we get from the ANCOVA model specification that

$$\begin{aligned}\begin{bmatrix}\widehat{\beta}_W \\ \widehat{\beta}_X\end{bmatrix} &= \left(\begin{bmatrix}\mathbb{W}^\top \\ \mathbb{X}^\top\end{bmatrix}\begin{bmatrix}\mathbb{W} & \mathbb{X}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbb{W}^\top \\ \mathbb{X}^\top\end{bmatrix}\mathbb{Y} = \begin{bmatrix}\mathbb{W}^\top\mathbb{W} & \mathbb{W}^\top\mathbb{X} \\ \mathbb{X}^\top\mathbb{W} & \mathbb{X}^\top\mathbb{X}\end{bmatrix}^{-1}\begin{bmatrix}\mathbb{W}^\top\mathbb{Y} \\ \mathbb{X}^\top\mathbb{Y}\end{bmatrix} \\ &= \begin{bmatrix}\mathbb{W}^\top\mathbb{W} & 0 \\ 0 & \mathbb{X}^\top\mathbb{X}\end{bmatrix}^{-1}\begin{bmatrix}\mathbb{W}^\top\mathbb{Y} \\ \mathbb{X}^\top\mathbb{Y}\end{bmatrix} = \begin{bmatrix}(\mathbb{W}^\top\mathbb{W})^{-1} & 0 \\ 0 & (\mathbb{X}^\top\mathbb{X})^{-1}\end{bmatrix}\begin{bmatrix}\mathbb{W}^\top\mathbb{Y} \\ \mathbb{X}^\top\mathbb{Y}\end{bmatrix} \\ &= \begin{bmatrix}(\mathbb{W}^\top\mathbb{W})^{-1}\mathbb{W}^\top\mathbb{Y} \\ (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{Y}\end{bmatrix}.\end{aligned} \tag{2.33}$$

We recognise the entries of the last vector as the respective β estimates when running a regression with only $\mathbb{W}$ and $\mathbb{X}$, respectively. Specifically, this means that we expect that the $\widehat{\beta}_W$ estimate will be the same if we choose to include $\mathbb{X}$ in the regressions as if we do not adjust for baseline covariates (as for the ANOVA model). Later, in section 5.3.1, we will show more formally that the estimators are consistent, and thus that this holds in the asymptotic case, also holds when the ANCOVA model includes interaction effects between treatment allocation and all baseline covariates.

However, as we will now show, we expect the variance to decrease by adjusting for baseline covariates when assuming a specific model. We will consider a model where W and X are not orthogonal but instead we have demeaned W and X, such that in the case of an RCT we expect them to be orthogonal. Afterwards this model is transformed with the projection orthogonal to the space spanned by the columns in $\mathbb{X}$. From the FWL theorem, we get for the orthogonal projection $M_{\widetilde{\mathbb{X}}} = I_n - P_{\widetilde{\mathbb{X}}} = I_n - \widetilde{\mathbb{X}}(\widetilde{\mathbb{X}}^\top\widetilde{\mathbb{X}})^{-1}\widetilde{\mathbb{X}}^\top$, that we get an equivalent way of obtaining $\widehat{\beta}_W$ by determining the ML estimates on

$$M_{\widetilde{\mathbb{X}}}\widetilde{\mathbb{Y}} = M_{\widetilde{\mathbb{X}}}\widetilde{\mathbb{W}}\beta_W + \varepsilon \tag{2.34}$$

Now, we can express the variance of the estimated treatment effect from the ANCOVA model as

$$\mathbb{V}\mathrm{ar}\left(\widehat{\beta}_W\right) = \sigma^2\left(\widetilde{\mathbb{W}}^\top M_{\widetilde{\mathbb{X}}}\widetilde{\mathbb{W}}\right)^{-1} \approx \sigma^2\left(\widetilde{\mathbb{W}}^\top\widetilde{\mathbb{W}}\right)^{-1} \tag{2.35}$$

where the approximate equality follows since we expect $\widetilde{\mathbb{X}}$ to be orthogonal to $\widetilde{\mathbb{W}}$. When $\widetilde{\mathbb{X}}$ and $\widetilde{\mathbb{W}}$ are exactly orthogonal, we have an exact equality. The variance expression in (2.35) is the same as if we had just used W as covariate, but now we have a smaller error variance σ^2 since X is also included in the model. This is due to the error variance in the ANCOVA model (2.21) being smaller than the error variance in the ANOVA model (2.18). From this, under exact orthogonality between $\widetilde{\mathbb{W}}$ and $\widetilde{\mathbb{X}}$, we can see that when adjusting for covariates X, the standard error of the treatment estimator becomes smaller as the explanatory ability of X on Y increases.

This heuristic derivation could lead to the conclusion that we should adjust for all potential prognostic baseline covariates, since these are independent of W and could decrease the variance. While introducing baseline covariates in X that are not orthogonal to W, we potentially inflate the factor $\left(\widetilde{\mathbb{W}}^\top M_{\widetilde{\mathbb{X}}}\widetilde{\mathbb{W}}\right)^{-1}$ in equation (2.35), this does not happen when X and W are orthogonal. However, adjusting for many potentially prognostic coviariates, we run the risk of also adjusting for covariates that are not related to Y. Adjusting for covariates not related to Y will in general cause only a modest inflation of the variance estimate of the treatment estimator, also due to a loss of degrees of freedom, which inflates the estimate $\widehat{\sigma}^2$. However, if such covariates show signs of correlation to Y due to chance, this mistakenly decreases the variance estimate. This indicates that problems with overfitting are present, leading to a misleadingly low variance estimate of the treatment estimate, which inflates the type I error.

We note that these conclusions are based on assumptions of the true data generating process, namely that it can be described by a normal linear model $Y_i = W_i\beta_W + X_i\beta_X + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. However, these assumptions are not guaranteed to hold in practice. As an example, as we will later show in theorem 5.3.6, there are cases in which an ATE estimator obtained from an adjusted ANCOVA model has larger variance than its unadjusted counterpart. Specifically, when $\pi_1 \neq \pi_0$ and the relation between the covariates and the outcome is not the same across the treatment and control arms (which could be the case when the treatment effect is heterogeneous), the variance of the adjusted estimator could be larger if interaction effects are not modelled. In

the case of a heterogeneous treatment effect, the model specified without interaction terms would have heteroskedastic error terms if $\pi_1 \neq \pi_0$. Heuristically, this is the case since parts of the mean value structure is not explained if an interaction term WX is not included in the model, meaning that the residual variance is larger for either the treatment or the control group, whichever has the smallest number of participants (since minimising the squared residuals to obtain the MLE entails that each observation is weighted equally). This heteroskedasticity means that the first expression in (2.35) is not in fact the true variance. In section 3.5, we will return to the question of how to account for this in practice. ◀

As the example shows, adjusting for prognostic covariates reduces the variance, but whereas adjusting for covariates not related to the outcome is not expected to mistakingly decrease the variance estimate, it can happen due to chance. This is a reason why considerations need to be taken in regard to covariate adjustment.

2.3.2 Regulatory Guidelines for Covariate Adjustment

There are several considerations to be made when adjusting for covariates in RCTs, and both EMA, The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) in cooperation with EMA, and FDA have published guidelines for baseline covariate adjustment in clinical trials [16, 17, 18]. Baseline covariates that are known to be highly prognostic, and are specified as such in the protocol prior to the recording of the baseline values, should be included as covariates in the primary analysis. In general, few covariates should be included in the primary analysis, and justification on the use of each covariate should be provided. Some baseline values should always be included:

- If some covariate is stratified upon in the randomisation to ensure balance of treatments across the covariate, the covariate should be adjusted for in the analysis, except the case where the stratification was done purely for an administrative reason so the covariate is not expected to be related to the outcome.
- In case of a continuous primary outcome, the baseline value of this covariate should be included as a covariate whether the primary outcome is defined as "raw outcome" or the "change from baseline".

Adjustment for baseline covariates should not be included in the primary analysis if this was not prespecified, even if baseline imbalances are observed. This is to avoid analysts engaging in "fishing expeditions", creating different covariate adjusted models, and then *"focus on the covariate model that best accentuates the estimate and/or statistical significance of the treatment difference"* [19, 20]. Generally, no covariates measured after the randomisation should be included, as these have potentially been affected by the treatment.

3 Sample Size Calculations

In this chapter, we will describe methods for sample size calculations for the ANOVA model z- and t-tests as well as for the ANCOVA model t-test. Furthermore, we will discuss how vulnerable the ANCOVA model is to violations of some of its assumptions on the true data generating process, and how to alleviate such potential violations. We conclude the chapter by considering corrections to the sample size calculations when multiple hypotheses need to be tested.

3.1 General Considerations

The overall goal when conducting sample size calculations is to determine the number of participants in a trial required to ensure a prespecified power $1-\beta$, where β is the probability of a type II error. The power is calculated dependent on the specific statistical test that will be applied at a significance level α. The basic elements of a sample size calculation is an $\mathcal{H}_0$-hypothesis and alternative $\mathcal{H}_1$-hypothesis involving a test statistic T and a rejection region R_α. The probability of type I error can then be expressed as

$$\mathbb{P}\left(T \in R_\alpha \mid \mathcal{H}_0\right) = \alpha. \tag{3.1}$$

In the case of the alternative hypothesis being true, the power can be expressed as

$$\mathbb{P}\left(T \in R_\alpha \mid \mathcal{H}_1\right) = 1-\beta. \tag{3.2}$$

In other words, the power is a measure of how likely we are to correctly reject the null hypothesis, using a specific test with a certain sample size. The aim of sample size calculations is then to determine the minimum sample size required such that equation (3.2) is fulfilled for some specified β while ensuring a type I error rate of α, according to equation (3.1).

We consider the case where T is continuous and the test problem is one-sided with large values being critical to $\mathcal{H}_0$, that is, the $\mathcal{H}_0$-hypothesis is rejected for sufficiently large values of T. In this situation we have $R_\alpha =]c_\alpha, \infty[$ for some critical value $c_\alpha \in \mathbb{R}$. We let F_0 and F_1 denote the cumulative distribution functions of T under the $\mathcal{H}_0$- and $\mathcal{H}_1$-hypothesis, respectively. We then have

$$\begin{aligned} F_0(c_\alpha) &= \mathbb{P}\left(T < c_\alpha \mid \mathcal{H}_0\right) = 1-\alpha \\ F_1(c_\alpha) &= \mathbb{P}\left(T < c_\alpha \mid \mathcal{H}_1\right) = \beta. \end{aligned} \tag{3.3}$$

Assuming that the inverses of F_0 and F_1 exist, we obtain

$$c_\alpha = F_0^{-1}(1-\alpha) = F_1^{-1}(\beta) \quad \Rightarrow \quad F_1\left(F_0^{-1}(1-\alpha)\right) = \beta. \tag{3.4}$$

From the right hand side of equation (3.4) we see the relation between the significance level, α, and the probability of making a type II error, β. If we fix the sample size n and then decrease α, then $1-\alpha$ moves closer to 1, meaning that $F_0^{-1}(1-\alpha)$ increases, implying that β increases. Thus, when decreasing the significance level and hence the probability of making a type I error, the price to be paid is in the form of decreasing power.

Usually F_1 depends on n, and therefore we can use equation (3.4) to determine n. By solving the equation for a prespecified β, we would often not obtain an integer n. Therefore we seek to obtain the smallest positive integer n such that $F_0(c_\alpha) = 1-\alpha$ while $F_1(c_\alpha) \leqslant \beta$, thereby maintaining a power of at least $1-\beta$. Since F_1 and its inverse are increasing functions, we then wish to obtain the smallest positive integer n such that

$$F_0^{-1}(1-\alpha) \leqslant F_1^{-1}(\beta) \quad \Rightarrow \quad F_1\left(F_0^{-1}(1-\alpha)\right) \leqslant \beta. \tag{3.5}$$

From these equations we see that we need to fix some desired α and β. Furthermore, we should know the distribution of the test statistic under $\mathcal{H}_0$, that is F_0, to be able to perform the test in equation (3.1). From F_0 and α, we can determine the critical value c_α required to perform this test from equation (3.3). As we will derive in the following sections, F_0 (and hence c_α) is not dependent on n for the ANOVA model z- and t-tests and the ANCOVA model t-test. Lastly, we regard F_1 as a function of n to determine the required n for the desired α and β [21, pp. 14–15]. We now provide an example which will prove to be useful when considering sample size calculations for the ANOVA model z- and t-tests and for the ANCOVA model t-test.

Example 3.1.1. F_0 **and** F_1 **being normal distributions with known standard deviation**
We assume that T is a test statistic satisfying $T \,|\, \mathcal{H}_0 \sim \mathcal{N}\left(0, \psi_0^2\right)$ and $T \,|\, \mathcal{H}_1 \sim \mathcal{N}\left(c(n), \psi_1^2\right)$ with known ψ_0^2 and ψ_1^2. This means that when the $\mathcal{H}_1$-hypothesis is true, the mean of the test statistic T is shifted by $c(n) > 0$. This implies that $T \,|\, \mathcal{H}_0 = Z\psi_0$ and $T \,|\, \mathcal{H}_1 = c(n) + Z\psi_1$ for $Z \sim \mathcal{N}(0,1)$, and we thus get

$$F_0^{-1}(1-\alpha) = z_{1-\alpha}\psi_0 \tag{3.6}$$

and

$$F_1^{-1}(\beta) = c(n) + z_\beta\psi_1 = c(n) - z_{1-\beta}\psi_1, \tag{3.7}$$

with z_p denoting the p'th quantile of the standard normal cumulative distribution function. Now using the first part of equation (3.4), we get

$$c(n) = z_{1-\alpha}\psi_0 + z_{1-\beta}\psi_1, \tag{3.8}$$

which should then be solved for n [21, p. 16]. ◀

The conducted hypotheses and corresponding tests should be chosen according to the aim of the clinical trial. Specifically, a trial can be a *superiority trial*, *non-inferiority trial* or *equivalence trial*. Denoting the true average treatment effect by $\Delta = \mathbb{E}\big[Y(1) - Y(0)\big]$, we assume without loss of generality that $\Delta > 0$ implies that the treatment group has a better outcome.

A trial is a superiority trial if the aim is to show that a treatment is superior to a control treatment. In such a trial, the null hypothesis and alternative hypothesis are

$$\mathcal{H}_0 : \ \Delta \leqslant \Delta_s \qquad \text{and} \qquad \mathcal{H}_1 : \ \Delta > \Delta_s, \tag{3.9}$$

where $\Delta_s \geqslant 0$ is a predetermined *superiority margin*. In a non-inferiority trial we seek to show that the treatment is not worse than the control by more than some prespecified amount. Therefore, the null hypothesis and alternative hypotheses are

$$\mathcal{H}_0 : \ \Delta \leqslant -\Delta_{ni} \qquad \text{and} \qquad \mathcal{H}_1 : \ \Delta > -\Delta_{ni}, \tag{3.10}$$

where $\Delta_{ni} \geqslant 0$ is a predetermined *non-inferiority margin*. Lastly, in an equivalence trial the aim is to show that $\Delta \in [-\Delta_e, \Delta_e]$, with $\Delta_e \geqslant 0$ being a prespecified *equivalence margin*. Therefore we have

$$\mathcal{H}_0 : \ |\Delta| - \Delta_e > 0 \qquad \text{and} \qquad \mathcal{H}_1 : \ |\Delta| - \Delta_e \leqslant 0. \tag{3.11}$$

In practice an equivalence trial is often not used, since the non-inferiority trial contains the equivalence hypothesis but enables the possibility for the treatment to show superiority. The values of Δ_s, Δ_{ni} and Δ_e depend on the specific context of the trial and scale of the outcome variable [22].

As described earlier, when drop-outs are present due to e.g. violation of the protocol or adverse events, we use the intention-to-treat principle. If data is incomplete, values should be imputed using a multiple imputation method. This would lead to a loss in power due to a loss in information, and hence withdrawals should be taken into account when determining the sample size. A method for doing so is to assume a withdrawal rate and then adjust the sample size conservatively. Specifically, if a sample size calculation gives a required sample size of n_{ss}, then we could instead recruit $n = n_{\text{ss}}/(1-q)$ participants, where q is the assumed dropout rate. If the dropout rate is expected to differ between the treatment and control group, this correction would be done groupwise for n_0 and n_1. This would imply that the power is not less than the desired value, since the method conservatively assumes that the withdrawals do not provide any information in regard to the primary outcome of the patients that dropped out [21, p. 10].

3.2 ANOVA Model z-test

We will now derive sample size formulas for normally distributed outcomes with known standard deviation in the context of a superiority trial, using an ANOVA model to estimate the ATE. Often σ_w^2 is not known in practice, which is an issue we will return to later. For now, we consider

the setup where the outcome variables $Y_{wj} \sim \mathcal{N}(\mu_w, \sigma_w^2)$ are mutually independent for $w = 0, 1$ and $j = 1, 2, \ldots, n_w(n)$. We note that this setup entails a violation of the assumption of homoskedasticity of the normal linear model, but as we will show, the distribution of the z-test statistic still holds when we account for this.

Under a superiority trial in which $\Delta > 0$ implies that the treatment group has a better outcome, we wish to test the hypotheses in equation (3.9). Thereby we have a one-sided test problem given by

$$\mathcal{H}_0 : \ \Delta - \Delta_s \leqslant 0 \qquad \text{and} \qquad \mathcal{H}_1 : \ \Delta - \Delta_s > 0. \tag{3.12}$$

To test this hypothesis we can use the z-test statistic given by

$$Z = \frac{\overline{Y}_1 - \overline{Y}_0 - \Delta_s}{\sqrt{\frac{1}{n_1(n)}\sigma_1^2 + \frac{1}{n_0(n)}\sigma_0^2}}, \tag{3.13}$$

where we estimate μ_w by the group mean $\overline{Y}_w = \frac{1}{n_w(n)} \sum_{j=1}^{n_w(n)} Y_{wj}$, corresponding to the ANOVA ML estimate seen in appendix A.1. Then $\overline{Y}_1 - \overline{Y}_0$ follows a normal distribution with

$$\mathbb{E}\left[\overline{Y}_1 - \overline{Y}_0\right] = \mu_1 - \mu_0 = \Delta \tag{3.14}$$

and

$$\mathbb{V}\text{ar}\left(\overline{Y}_1 - \overline{Y}_0\right) = \frac{1}{n_1(n)}\sigma_1^2 + \frac{1}{n_0(n)}\sigma_0^2. \tag{3.15}$$

This implies that

$$Z \sim \mathcal{N}\left(\frac{\mu_1 - \mu_0 - \Delta_s}{\sqrt{\frac{1}{n_1(n)}\sigma_1^2 + \frac{1}{n_0(n)}\sigma_0^2}},\ 1\right). \tag{3.16}$$

The null hypothesis includes the situation where $\Delta - \Delta_s = 0$, in which case the mean value of the Z statistic is 0. In the remaining situations of $\Delta - \Delta_s < 0$, the mean value of the Z statistic is smaller than 0 with the same variance. That is, if $\Delta - \Delta_s = 0$, we get the largest critical value $z_{1-\alpha}$, meaning that R_α is as small as possible. Since we wish to reject all possible cases of $\mathcal{H}_0$ being true, we look at the situation where $\Delta - \Delta_s = 0$, entailing that $Z \sim \mathcal{N}(0, 1)$. The rejection region is then $R_\alpha = \{z \mid z > z_{1-\alpha}\}$ for a significance level α, since these values are critical to $\mathcal{H}_0$. Therefore we can use example 3.1.1 to write the mean parameter of the z-test statistic under $\mathcal{H}_1$ as

$$c(n) = \frac{\mu_1 - \mu_0 - \Delta_s}{\sqrt{\frac{1}{n_1(n)}\sigma_1^2 + \frac{1}{n_0(n)}\sigma_0^2}} = z_{1-\alpha} + z_{1-\beta}, \tag{3.17}$$

which should be solved for n. If we assume equally sized groups, we obtain

$$\begin{aligned} \frac{\mu_1 - \mu_0 - \Delta_s}{\sqrt{\frac{1}{n_1(n)}(\sigma_1^2 + \sigma_0^2)}} &= z_{1-\alpha} + z_{1-\beta} \\ \Rightarrow \quad \frac{\mu_1 - \mu_0 - \Delta_s}{z_{1-\alpha} + z_{1-\beta}} &= \sqrt{\frac{1}{n_1(n)}(\sigma_1^2 + \sigma_0^2)} \\ \Rightarrow \quad \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\mu_1 - \mu_0 - \Delta_s}\sqrt{\sigma_1^2 + \sigma_0^2}\right)^2 &= n_1(n) = \frac{n}{2}. \end{aligned} \tag{3.18}$$

If we instead assume homoskedasticity, that is, $\sigma_1 = \sigma_0 = \sigma_Y$, and let the ratio of the groups sizes be $r = \frac{n_1(n)}{n_0(n)}$, we have

$$\begin{aligned} z_{1-\alpha} + z_{1-\beta} = \frac{\mu_1 - \mu_0 - \Delta_s}{\sqrt{\frac{n_0(n)+n_1(n)}{n_1(n)n_0(n)}\sigma_Y^2}} &= \sqrt{\frac{n_1(n)n_0(n)}{n_0(n)+n_1(n)}}\frac{\mu_1 - \mu_0 - \Delta_s}{\sigma_Y} \\ &= \sqrt{\frac{r}{(1+r)^2}\left(n_0(n)+n_1(n)\right)}\frac{\mu_1 - \mu_0 - \Delta_s}{\sigma_Y}, \end{aligned} \tag{3.19}$$

using that

$$\begin{aligned} \frac{n_1(n)n_0(n)}{n_0(n)+n_1(n)} &= \frac{n_1(n)}{n_0(n)}\left(\frac{n_0(n)}{n_0(n)+n_1(n)}\right)^2\left(n_0(n)+n_1(n)\right) \\ &= r\left(\frac{1}{\frac{n_0(n)+n_1(n)}{n_0(n)}}\right)^2\left(n_0(n)+n_1(n)\right) \\ &= \frac{r}{(1+r)^2}\left(n_0(n)+n_1(n)\right). \end{aligned} \tag{3.20}$$

Therefore we have

$$n = n_0(n) + n_1(n) = \frac{(1+r)^2}{r}(z_{1-\alpha} + z_{1-\beta})^2\left(\frac{\sigma_Y}{\mu_1 - \mu_0 - \Delta_s}\right)^2. \tag{3.21}$$

In general, the group wise sample sizes are given in terms of r by the relations

$$n_1(n) = \left(\frac{r}{1+r}\right)n \qquad \text{and} \qquad n_0(n) = \left(\frac{1}{1+r}\right)n. \tag{3.22}$$

In practice, trying to determine a sample size from a desired allocation ratio could lead to non-integer values of the group sample sizes, which should then be rounded up to ensure a power of at least $1-\beta$. Such a rounding could lead to slight differences in the allocation ratio r. Another

alternative when r is an integer would be to round to the smallest positive integer which $r+1$ divides. For example if $r = 3$ and we have obtained $n = 162.4$ we would obtain a sample size of $n = 164$. Thereby, we have the same allocation ratio as before. If we had instead rounded each group we would have $162.4 \cdot 3/4 = 121.8$ rounded to 122 and $162.4 \cdot 1/4 = 40.6$ rounded to 41. This gives an allocation ratio of $122/41 = 2.98$.

Instead of finding the sample size from a fixed allocation ratio, we can investigate which allocation ratio is optimal in regard to minimising the required sample size, which we do in the following example.

Example 3.2.1. Required sample size as a function of the allocation ratio
We wish to investigate the impact of the allocation ratio on the required n in the situation of $\sigma_1 = \sigma_0 = \sigma_Y$. Specifically, we wish to obtain the most optimal allocation ratio such that the required n is as small as possible. For that purpose, we will regard the required n in the representation (3.21) as a function of the allocation ratio $r > 0$. Disregarding the constant factor, the first and second order derivatives are given as

$$\begin{aligned} n'(r) &= \frac{2(1+r)}{r} - \frac{(1+r)^2}{r^2} = \frac{r^2-1}{r^2} = 1 - \frac{1}{r^2} \\ n''(r) &= \frac{2r}{r^2} - \frac{2(r^2-1)}{r^3} = \frac{2}{r^3} > 0, \end{aligned} \tag{3.23}$$

so $n(r)$ is minimised when $r = 1$ meaning that we have equally sized groups.

The percentage-wise increase of the total required sample size compared to balanced allocation is given by

$$\begin{aligned} \frac{n(r) - n(1)}{n(1)} &= \frac{\left(\frac{(1+r)^2}{r} - 4\right)(z_{1-\alpha} + z_{1-\beta})^2 \left(\frac{\sigma_Y}{\mu_1 - \mu_0 - \Delta_s}\right)^2}{4(z_{1-\alpha} + z_{1-\beta})^2 \left(\frac{\sigma_Y}{\mu_1 - \mu_0 - \Delta_s}\right)^2} \\ &= \frac{(1+r)^2}{4r} - 1 = \frac{1 + r^2 + 2r - 4r}{4r} = \frac{(r-1)^2}{4r}. \end{aligned} \tag{3.24}$$

From this we see that for imbalances such as $r = 3/2$ or $r = 2$ we have a 4.16% and 12.5% increase in the required sample size, respectively. For more extreme imbalances such as $r = 4$ and $r = 6$ we have a 56.25% and 104.16% increase, respectively. ◀

Example 3.2.1 shows that an allocation ratio of 1 always provides the smallest required sample size when we wish to investigate the efficacy of a drug. However, considerations regarding ethics, the power to detect adverse events in the treatment group, the expected rate of dropouts and the economic cost of the study could require a non-balanced allocation ratio [23].

If we had instead considered a non-inferiority trial we would have obtained similar results just substituting $-\Delta_s$ with Δ_{ni} [21, pp. 19–21].

From equation (3.21) we see what in general is needed to determine a sample size that ensures a type I error rate of α and a desired power of $1-\beta$. The significance level α is needed, and we see that as α decreases, $z_{1-\alpha}$ increases, yielding an increase in the required n. In the ICH E9 Guideline [17, p. 27], it is stated that: *"The approach of setting type I errors for one-sided tests at half the conventional type I error used in two-sided tests is preferable in regulatory settings."* This ensures consistency between the confidence intervals for one-sided tests and the corresponding two-sided tests. In practice this means that we should require the same sample size regardless of whether we conduct a one-sided or two-sided test. Thus, in practice a one-sided test should be conducted with a significance level of $\alpha/2$. Furthermore, we need to predetermine the desired power $1-\beta$ of the analysis. Again we see that if β decreases, the required n increases. We also need an assumed effect size $\mu_1 - \mu_0$ together with the superiority or non-inferiority margin. The smaller the assumed effect size is, the larger n is needed to have a $1-\beta$ probability of determining this effect. Increasing the margin also increases the required n. Lastly, we need the allocation ratio r, which we have already shown leads to the smallest n when we have equally sized groups.

In general we want a low probability of conducting a type I error, since such an error would lead to approval of a potentially non-effective or worse drug. In medical research a two-sided significance level of $\alpha = 0.05$ is the standard approach [21, p. 8]. Now, replacing α with $\alpha/2$ in equation (3.21) according to the ICH Guidelines, we can determine the cost of more power in terms of required increase in n. For example, having predetermined $\alpha/2 = 0.025$, $\mu_1 - \mu_0 = 10$, $\Delta_s = 5$, $\sigma_Y = 20$ and $r = 1$ we can determine that going from 0.8 to 0.9 in power would yield an increase in the required n by a factor 1.34. In general, there is a tendency of a large increase in the required sample size when increasing the desired power from 80% [21, p. 22]. Because of such a large increase in required sample size, a larger probability of type II errors are generally accepted than for the type I error. In general, clinical trials usually have a power of at least 80% [24]. On the other hand, if the trial cannot be repeated, a low β is needed, since a type II error in this case could lead to never determining the potential of the drug [21, p. 8].

Equation (3.18) shows a sample size formula in case of equally sized groups, and (3.21) shows what is needed when we assume constant variance across the treatment- and control groups. Specifically, we see that the calculation is based on known variance, with larger variance implying a larger required sample size; in order to maintain certainty in the estimate when data varies a lot, we need more participants. The variance is called a *nuisance parameter*, since it is not a part of the hypothesis under investigation but still is a part of the calculation of power. In practice the variance can be estimated by the previously conducted phase I or II trials, the preclinical trial or by some other a priori knowledge. The a priori knowledge can be obtained from a meta-analysis of a collection of previous trials. Since there is uncertainty in regard to the estimated nuisance parameter, a sensitivity analysis should be conducted, where the sample size is calculated for multiple values of the parameter. It has been shown analytically that the probability of obtaining the planned power is less than 50% when using a sample variance estimate for the z-test. This illustrates the importance of a sensitivity analysis. It has also been shown analytically that if we use the upper boundary of a $(1-\gamma)$-confidence interval for an estimate $\widehat{\sigma}^2_Y$ in place of σ^2_Y, then there is a probability of $1-\gamma$ of gaining a power of $1-\beta$ for the z-test [21, p. 27] [25].

3.3 ANOVA Model t-test

We consider the same setup as in section 3.2, but now with unknown variance $\sigma_Y^2 = \mathbb{V}\mathrm{ar}(Y(w))$ for $w = 0, 1$, which is assumed to be the same across the treatment- and control groups. Again, we start by considering the superiority trial with the same hypotheses as in section 3.2.

We denote the standard ANOVA pooled sample standard deviation as

$$S_Y = \sqrt{\frac{\left(n_1(n) - 1\right) S_1^2 + \left(n_0(n) - 1\right) S_0^2}{n_1(n) + n_0(n) - 2}} \tag{3.25}$$

where the estimates of the group variances are

$$S_w^2 = \frac{1}{n_w(n) - 1} \sum_{j=1}^{n_w(n)} \left(Y_{wj} - \overline{Y}_w\right)^2, \qquad w = 0, 1. \tag{3.26}$$

We note that S_Y^2 is the usual unbiased variance estimator $\widehat{\sigma}_Y^2$ for this normal linear model, since a subject in group w is predicted by the group mean $\overline{Y}_w$. This is also described in appendix A.2.

The ANOVA model t-test statistic is then defined by

$$T_{\text{ANOVA}} = \frac{\overline{Y}_1 - \overline{Y}_0 - \Delta_s}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}}\left(\overline{Y}_1 - \overline{Y}_0\right)}} = \frac{\sqrt{\frac{n_1(n)n_0(n)}{n_1(n)+n_0(n)}}\left(\overline{Y}_1 - \overline{Y}_0 - \Delta_s\right)}{S_Y}, \tag{3.27}$$

where

$$\begin{aligned} \mathbb{V}\mathrm{ar}\left(\overline{Y}_1 - \overline{Y}_0\right) = \mathbb{V}\mathrm{ar}\left(\overline{Y}_1\right) + \mathbb{V}\mathrm{ar}\left(\overline{Y}_0\right) &= \left(\frac{1}{n_1(n)} + \frac{1}{n_0(n)}\right)\sigma_Y^2 \\ &= \frac{n_1(n) + n_0(n)}{n_1(n)n_0(n)}\sigma_Y^2 \end{aligned} \tag{3.28}$$

is estimated by

$$\widehat{\mathbb{V}\mathrm{ar}}\left(\overline{Y}_1 - \overline{Y}_0\right) = \frac{n_1(n) + n_0(n)}{n_1(n)n_0(n)} S_Y^2. \tag{3.29}$$

Since S_Y^2 is the standard variance estimator from a normal linear model, we have that

$$S_Y^2 \sim \frac{\sigma_Y^2}{n_1(n) + n_0(n) - 2}\chi^2\left(n_1(n) + n_0(n) - 2\right), \tag{3.30}$$

with $\overline{Y}_1$ and $\overline{Y}_0$ being independent of S_Y^2, making the numerator and denominator in equation (3.27) independent. Notice that from the distribution, we can conclude that S_Y^2 is an unbiased estimate of σ_Y^2, since the expected value of the scaled χ^2-distribution on the right hand side is 1.

As in section 3.2 with $\Delta - \Delta_s = 0$ being the situation with R_α as small as possible in the case of $\mathcal{H}_0$ being true, we have

$$\frac{\overline{Y}_1 - \overline{Y}_0 - \Delta_s}{\sqrt{\left(\frac{1}{n_1(n)} + \frac{1}{n_0(n)}\right)\sigma_Y^2}} = \frac{\sqrt{\frac{n_1(n)n_0(n)}{n_1(n)+n_0(n)}}\left(\overline{Y}_1 - \overline{Y}_0 - \Delta_s\right)}{\sigma_Y} \sim \mathcal{N}(0,1), \tag{3.31}$$

under the $\mathcal{H}_0$-hypothesis.

From independence of the numerator and denominator of the last expression in equation (3.27), using the definition of the central Student's t-distribution together with (3.31) and (3.30), the t-statistic has the distribution

$$T_{\text{ANOVA}} = \frac{\sqrt{\frac{n_1(n)n_0(n)}{n_1(n)+n_0(n)}}\left(\overline{Y}_1 - \overline{Y}_0 - \Delta_s\right) \Big/ \sigma_Y}{\sqrt{S_Y^2/\sigma_Y^2}} \sim t\left(n_1(n) + n_0(n) - 2\right) \tag{3.32}$$

under the $\mathcal{H}_0$-hypothesis.

For the one-sided test, the rejection region is $R_\alpha = \left\{t \;\middle|\; t > t_{1-\alpha/2,\, n_1(n)+n_0(n)-2}\right\}$ using the ICH E9 guideline for the significance level. Under $\mathcal{H}_1$ we use the definition of the non-central Student's t-distribution, where the non-centrality parameter is defined as the mean of the numerator in (3.32). This numerator coincides with the z-test statistic in (3.13), which has the mean specified in (3.16). That is, the non-centrality parameter can be expressed as

$$c(n) = \sqrt{\frac{r}{(1+r)^2} \cdot n} \cdot \frac{\mu_1 - \mu_0 - \Delta_s}{\sigma_Y}, \tag{3.33}$$

using equation (3.19). Thus, $c(n)$ again depends on σ_Y, and in practice one can use an estimated variance. As with the z-test, a sensitivity analysis should therefore be conducted using different estimates for σ_Y.

According to equation (3.5) we can use the cumulative distribution functions for the t-distribution $F_0 = F_{t,n-2,0}$ and $F_1 = F_{t,n-2,c(n)}$ with $n-2$ degrees of freedom, and non-centrality parameter equal to 0 and $c(n)$, respectively, to determine the smallest positive integer n such that

$$F_{t,n-2,0}^{-1}(1-\alpha/2) \leqslant F_{t,n-2,c(n)}^{-1}(\beta). \tag{3.34}$$

The equation can be solved iteratively by inserting an n_k on the right hand side for the k'th iteration and then determining the smallest integer n_{k+1} such that the inequality holds, using

n_{k+1} on the left hand side. This n_{k+1} is then used to evaluate the right hand side, starting the next iteration. The iterations continue until $n_k = n_{k+1}$. A good starting point could be the n found by conducting a z-test assuming a specific value of σ_Y. As for the z-test, it can be shown that an allocation ratio of $r = 1$ results in the smallest required n. An alternative to iteratively solving for n, the solution can be approximated as described in the next section.

3.3.1 Approximation Formulas

Different approximations for the solution of equation (3.34) have been proposed. In practice it is no problem to use the iterative method, but to illustrate that a larger sample is needed for the t-test, we will introduce an approximate solution. An approximation that has proven to be good in practice is the Guenther and Schouten approximation [26, 27] [21, p. 25], which is given by

$$n_{\text{GS},0}(n) = \frac{1+r}{r}(z_{1-\alpha/2} + z_{1-\beta})^2 \left(\frac{\sigma_Y}{\mu_1 - \mu_0 - \Delta_s}\right)^2 + \frac{(z_{1-\alpha/2})^2}{2(1+r)} \tag{3.35}$$

and

$$n_{\text{GS},1}(n) = r n_{\text{GS},0}(n), \tag{3.36}$$

where σ_Y is replaced by an estimate in practice, and we can again either round the total sample size or each group sample to obtain integer values. The first term in equation (3.35) is motivated by the t-distribution being approximately normal for a large n. We could then use that $n = (1 + r)n_0$ in equation (3.21) to obtain the first term. The last term is a correction, since the t-distribution is only approximately normal. Thus, we have

$$\begin{aligned} n_{\text{GS}} &= n_{\text{GS},0}(n) + n_{\text{GS},1}(n) \\ &= (1+r)\left(\frac{1+r}{r}(z_{1-\alpha/2} + z_{1-\beta})^2 \left(\frac{\sigma_Y}{\mu_1 - \mu_0 - \Delta_s}\right)^2 + \frac{(z_{1-\alpha/2})^2}{2(1+r)}\right) \\ &= n_{z\text{-test}} + \frac{(z_{1-\alpha/2})^2}{2}. \end{aligned} \tag{3.37}$$

Therefore if $\alpha = 0.05$ we have $\frac{(z_{1-\alpha/2})^2}{2} = 1.92$ meaning that the required sample size increases by 2 when using the t-test instead of the z-test.

Again, if we instead consider a non-inferiority trial, we can obtain similar results just substituting Δ_s with $-\Delta_{ni}$.

All considerations at the end of section 3.2 also apply to this case of the t-test, where choices have to be made about significance level, power, assumed effect size, margin and nuisance parameter when wishing to determine a sample size, whether we wish to solve iteratively or approximately.

3.4 ANCOVA Model t-test

We will now derive sample size formulas for the ANCOVA model described in section 2.3. We start by considering the case of adjustment for one covariate, and we will then generalise to the case of adjusting for several covariates. The results are mostly based on [28] and [25], in which most equations are not explicitly derived. Where this is the case, we carry out the calculations, the most comprehensive of which are placed in appendices A.3 and A.4.

3.4.1 Univariable Covariate Adjustment

We first consider the case of adjusting for one random covariate X_{wj}, with $w = 1, 0$ indicating the treatment- and control group, respectively, and $j = 1, 2, \ldots, n_w$ indicating the patient number in the corresponding group, and equivalently for the response Y_{wj}. To determine the MLE in the ANCOVA model (2.20), we can write an equivalent reparametrised form of the ANCOVA model as

$$Y_{wj} = (1-w)\beta_0 + w\beta_1 + X_{wj}\beta_X + \varepsilon_{wj}, \qquad j = 1, 2, \ldots, n_w, \quad w = 0, 1, \tag{3.38}$$

where $\varepsilon = [\varepsilon_{11}, \varepsilon_{12}, \ldots, \varepsilon_{0n_0}]^\top \sim \mathcal{N}\left(0, \mathrm{diag}_{n_1+n_0}(\sigma^2, \sigma^2, \ldots, \sigma^2)\right)$ and we again assume that the X_{wj}'s are mutually independent, the Y_{wj}'s are mutually independent, and all the X_{wj}'s and Y_{wj}'s are independent of ε. We denote the design matrix of this model as $\mathbb{D}$, which is explicitly defined in appendix A.3. Furthermore, using the assumption of homoskedasticity, we denote as σ_Y^2 the variance of $Y(w)$ for $w = 0, 1$, by σ_X^2 the variance of the covariate, and the correlation as $\rho = \sigma_{XY}/(\sigma_X\sigma_Y)$, for σ_{XY} denoting the covariance between X and Y. From this model specification, we get $\beta_1 - \beta_0 = \mathrm{CATE} = \mathrm{ATE}$, as seen in section 2.3.

In the case of a superiority trial, the $\mathcal{H}_0$ and $\mathcal{H}_1$-hypotheses can then be formulated according to the new parametrisation as

$$\mathcal{H}_0 : (\beta_1 - \beta_0) - \Delta_s \leqslant 0 \qquad \text{and} \qquad \mathcal{H}_1 : (\beta_1 - \beta_0) - \Delta_s > 0. \tag{3.39}$$

In appendix A.3, the ML estimate for the control and treatment group coefficients are derived as

$$\widehat{\beta}_k = -\overline{X}_k\widehat{\beta}_X + \overline{Y}_k, \qquad k = 0, 1, \tag{3.40}$$

using the estimator of β_X in equation (A.12). This gives an unbiased ATE estimator as $\widehat{\beta}_1 - \widehat{\beta}_0$. The variance of the ATE estimator is derived in appendix A.4 as

$$\mathbb{V}\mathrm{ar}\left(\widehat{\beta}_1 - \widehat{\beta}_0\right) = \left(\frac{\left(\overline{X}_1 - \overline{X}_0\right)^2}{S_X^2(n-2)} + n_1^{-1} + n_0^{-1} \right)\sigma^2, \tag{3.41}$$

where

$$S_X = \sqrt{\frac{\sum_{w=0}^{1}\sum_{j=1}^{n_w}\left(X_{wj} - \overline{X}_w\right)^2}{n-2}} \tag{3.42}$$

is a pooled variance estimator of the covariate. In appendix A.4 an unbiased estimate of the variance is derived as

$$\begin{aligned}\widehat{\mathbb{V}\mathrm{ar}}\left(\widehat{\beta}_1 - \widehat{\beta}_0\right) &= \left(\frac{\left(\overline{X}_1 - \overline{X}_0\right)^2}{S_X^2(n-2)} + n_1^{-1} + n_0^{-1}\right)\widehat{\sigma}^2 \\ &= \left(\frac{\left(\overline{X}_1 - \overline{X}_0\right)^2}{S_X^2(n-2)} + n_1^{-1} + n_0^{-1}\right)\frac{n-2}{n-3}S_Y^2\left(1-\widehat{\rho}^2\right),\end{aligned} \tag{3.43}$$

using the usual unbiased estimator of σ^2 given as

$$\widehat{\sigma}^2 = \frac{(\mathbb{Y} - \mathbb{D}\widehat{\beta})^\top(\mathbb{Y} - \mathbb{D}\widehat{\beta})}{n-3} \sim \frac{\sigma^2}{n-3}\chi^2(n-3), \tag{3.44}$$

and a pooled estimate of the correlation between X and Y is given as

$$\widehat{\rho} = \frac{\sum_{w=0}^{1}\sum_{j=1}^{n_w}\left(X_{wj} - \overline{X}_w\right)\left(Y_{wj} - \overline{Y}_w\right)}{\sqrt{\sum_{w=0}^{1}\sum_{j=1}^{n_w}\left(X_{wj} - \overline{X}_w\right)^2\sum_{w=0}^{1}\sum_{j=1}^{n_w}\left(Y_{wj} - \overline{Y}_w\right)^2}}. \tag{3.45}$$

Similar to section 3.3, the t-test statistic is given by

$$T_{\text{ANCOVA}} = \frac{\widehat{\beta}_1 - \widehat{\beta}_0 - \Delta_s}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}}\left(\widehat{\beta}_1 - \widehat{\beta}_0\right)}}. \tag{3.46}$$

As in section 3.2 with $\widehat{\beta}_1 - \widehat{\beta}_0 - \Delta_s = 0$ being the situation with R_α as small as possible in the case of $\mathcal{H}_0$ being true, we have that the numerator of the t-test statistic divided by the true variance has the distribution

$$\frac{\widehat{\beta}_1 - \widehat{\beta}_0 - \Delta_s}{\sqrt{\sigma^2\left(\frac{(\overline{X}_1 - \overline{X}_0)^2}{S_X^2(n-2)} + n_1^{-1} + n_0^{-1}\right)}} \sim \mathcal{N}(0,1), \tag{3.47}$$

under the $\mathcal{H}_0$-hypothesis. Therefore under the $\mathcal{H}_0$-hypothesis and using the definition of the central Student's t-distribution with $\widehat{\sigma}^2$ and $\widehat{\beta}$ being independent, we have from (3.43), (3.44) and (3.47), that

$$\begin{aligned} T_{\text{ANCOVA}} &= \frac{\left(\widehat{\beta}_1 - \widehat{\beta}_0 - \Delta_s\right) \Big/ \sqrt{\sigma^2 \left(\frac{(\overline{X}_1 - \overline{X}_0)^2}{S_X^2(n-2)} + n_1^{-1} + n_0^{-1}\right)}}{\sqrt{\widehat{\mathbb{V}\text{ar}}\left(\widehat{\beta}_1 - \widehat{\beta}_0\right) \Big/ \sigma^2 \left(\frac{(\overline{X}_1 - \overline{X}_0)^2}{S_X^2(n-2)} + n_1^{-1} + n_0^{-1}\right)}} \\ &= \frac{\left(\widehat{\beta}_1 - \widehat{\beta}_0 - \Delta_s\right) \Big/ \sqrt{\sigma^2 \left(\frac{(\overline{X}_1 - \overline{X}_0)^2}{S_X^2(n-2)} + n_1^{-1} + n_0^{-1}\right)}}{\sqrt{\widehat{\sigma}^2 \Big/ \sigma^2}} \sim t(n-3). \end{aligned} \tag{3.48}$$

This is again a one-sided test, and the rejection region is $R_\alpha = \left\{t \;\middle|\; t > t_{1-\alpha/2,n-3}\right\}$, where we again use the ICH E9 guideline for the significance level [21, pp. 29–30].

Under the $\mathcal{H}_1$-hypothesis we can use the definition of the non-central Student's t-distribution, where the non-centrality parameter is the expectation of the numerator of the ANCOVA t-test statistic (3.46) divided by the true variance $\mathbb{V}\text{ar}\left(\widehat{\beta}_1 - \widehat{\beta}_0\right)$. According to Kieser [21, p. 30], the non-centrality parameter can be expressed as

$$c(n) = \mathbb{E}\left[\frac{\widehat{\beta}_1 - \widehat{\beta}_0 - \Delta_s}{\sqrt{\sigma^2 \left(\frac{(\overline{X}_1 - \overline{X}_0)^2}{S_X^2(n-2)} + n_1^{-1} + n_0^{-1}\right)}}\right] = \sqrt{\frac{rn}{(1+r)^2}} \frac{\beta_1 - \beta_0 - \Delta_s}{\sigma_Y \sqrt{1-\rho^2}}. \tag{3.49}$$

Later, in equation (5.126), we will arrive at this conclusion in the asymptotic case. We note that the non-centrality parameter only differs from the non-centrality parameter for the t-test given in equation (3.33) by a factor $1/\sqrt{1-\rho^2}$. This implies that the two coincides for $\rho = 0$. Thus, when the covariate that we adjust for is uncorrelated to the outcome variable Y, the test reduces to a t-test in an ANOVA model. This is consistent with the intuition that it only makes sense to adjust for covariates that are related to the outcome. In the opposite case of ρ being numerically large, $\sqrt{1-\rho^2}$ decreases, so $c(n)$ becomes large, making it more likely to reject the null hypothesis in the case of $\mathcal{H}_1$ being true, thus gaining more power for a fixed sample size of n. This is consistent with the intuition that when X explains more of the variation in Y, we can be more certain that differences in the treatment- and control groups can be ascribed to treatment allocation.

We will show that this intuition translates directly to the t-test statistics of the ANCOVA and ANOVA models. We begin by considering the last expression of the variance estimate in equation (3.43). For sample sizes in clinical trials, we often have sufficiently large n, so that $\frac{n-2}{n-3} \approx 1$.

Furthermore, since the patients are randomly allocated to the treatment- and control groups, the covariates are equally distributed between the two groups in expectation, as seen in section 2.2.1. This, and n being sufficiently large, implies that $\frac{(\overline{X}_1-\overline{X}_0)^2}{S_X^2(n-2)}$ is small for large n. Therefore, we neglect this term in the variance estimate, and we get

$$\widehat{\mathbb{V}\mathrm{ar}}\left(\widehat{\beta}_1-\widehat{\beta}_0\right) \approx \left(1-\widehat{\rho}^2\right) S_Y^2 \left(n_1^{-1}+n_0^{-1}\right). \tag{3.50}$$

We can then approximate the ANCOVA t-test statistic (3.46) by

$$T_{\mathrm{ANCOVA}} \approx \frac{\widehat{\beta}_1-\widehat{\beta}_0-\Delta_s}{\sqrt{\left(1-\widehat{\rho}^2\right) S_Y^2 \left(n_1^{-1}+n_0^{-1}\right)}}. \tag{3.51}$$

Thus, this test statistic corresponds to the ANOVA test statistic in equation (3.32) but multiplied by the factor $1/\sqrt{1-\widehat{\rho}^2}$.

Using the non-centrality parameter in (3.49), we can use equation (3.5) where we use the cumulative distribution functions for the t-distribution $F_0 = F_{t,n-3,0}$ and $F_1 = F_{t,n-3,c(n)}$ to determine the smallest positive integer n such that

$$F^{-1}_{t,n-3,0}(1-\alpha/2) \leqslant F^{-1}_{t,n-3,c(n)}(\beta). \tag{3.52}$$

To obtain the minimum required sample size, we should again solve this iteratively by using the same procedure as for the ANOVA model t-test, discussed after equation (3.34). Furthermore, as for the ANOVA model t- and z-tests, it can be shown that the optimal allocation ratio is $r = 1$ in regard to minimising the required sample size. Alternatively, one can approximate the required sample size as described in the following section.

3.4.2 Approximation Formulas

In the following, we will derive some closed form approximation formulas for the required sample size. A central t-distribution converges towards a standard normal distribution implying that asymptotically, $T_{\mathrm{ANCOVA}} \xrightarrow{d} \mathcal{N}(0,1)$ under the $\mathcal{H}_0$-hypothesis. Similarly we obtain that asymptotically $T_{\mathrm{ANCOVA}} - c(n) \xrightarrow{d} \mathcal{N}(0,1)$ under the $\mathcal{H}_1$-hypothesis. Thus, we can use example 3.1.1 to approximately express the non-centrality parameter in equation (3.49) as

$$\sqrt{\frac{rn}{(1+r)^2}}\frac{\beta_1-\beta_0-\Delta_s}{\sigma_Y\sqrt{1-\rho^2}} \approx z_{1-\alpha/2}+z_{1-\beta}, \tag{3.53}$$

which should be solved for n. This results in the Frison-Pocock approximation formula [29] [21, p. 31], which is given by

$$n_{\mathrm{FP}} = \frac{(1+r)^2}{r}(z_{1-\alpha/2}+z_{1-\beta})^2\frac{\sigma_Y^2(1-\rho^2)}{(\beta_1-\beta_0-\Delta_s)^2}. \tag{3.54}$$

As in section 3.3, the Guenther-Schouten correction can now be used to adjust the sample size when approximating a t-distribution by a normal distribution. That is, we can approximate the required sample size by

$$n_{\mathrm{GS}} = n_{\mathrm{FP}} + \frac{(z_{1-\alpha/2})^2}{2}. \tag{3.55}$$

Again, using a significance level of $\alpha = 0.05$, the required sample size would be increased by 2. To determine n_1 and n_0, we can use equation (3.22) and either round the total sample size to an integer, or round each group sample to obtain integer values. In practice this approximation is very close to the exact sample size obtained from solving equation (3.52) for the smallest possible n.

From the approximation formulas derived in equations (3.54) and (3.55), we see that the sample size can be considerably decreased by using an ANCOVA model instead of an ANOVA model, since we in this case multiply the required sample size by a factor of $1 - \rho^2$. For instance if $\rho = 0.5$, the n found in equation (3.21), coming from the z-test, is multiplied by $(1 - 0.5^2) = 0.75$, thereby reducing the required sample size by 25% compared to the ANOVA model. For the t-test, the sample size calculations coming from the ANOVA and ANCOVA model are corrected from the z-test by the same negligible term $\frac{(z_{1-\alpha/2})^2}{2}$, and thus this calculated relative decrease in sample size, using ANCOVA over ANOVA, also holds almost exactly for the t-test. If the correlation is 0.7, the sample size is reduced by almost 50%. It is not unrealistic that the covariate that we adjust for is correlated to the outcome by a large magnitude. For example if we want to evaluate the end-of-treatment value or change from baseline in respect to some outcome, then we should adjust by the baseline value of the outcome variable, which is in accordance with regulatory guidelines as described in section 2.3.2. The baseline value is typically highly correlated to the outcome variable [21, pp. 30–32]. Thereby we obtain a substantial decrease of the required sample size.

In the case of a non-inferiority trial, we could obtain similar results substituting $-\Delta_s$ by Δ_{ni}.

All considerations at the end of section 3.2 also apply to this case of the ANCOVA model, where choices have to be made about significance level, power, assumed effect size, margin and nuisance parameters when wishing to determine a sample size, whether we wish to solve iteratively or approximately. In this case, both the variance and correlation coefficient are nuisance parameters, and it can be necessary to perform sensitivity analysis in regard to both parameters.

3.4.3 Multivariable Covariate Adjustment

We will now consider the case where we adjust by multiple covariates. The section is based on [30]. In this case we let $(Y_{wj}, X_{wj}) = (Y_{wj}, X^1_{wj}, X^2_{wj}, \ldots, X^p_{wj})$ be random independent observations for $w = 0, 1$ and $j = 1, \ldots, n_w$. Again, we consider an equivalent form of the

ANCOVA model parametrised as

$$Y_{wj} = (1-w)\beta_0 + w\beta_1 + \sum_{k=1}^{p} X_{wj}^k \beta_{X^k} + \varepsilon_{wj}, \qquad j = 1, 2, \ldots, n_w, \ w = 0, 1, \tag{3.56}$$

where $\varepsilon = [\varepsilon_{11}, \varepsilon_{12}, \ldots, \varepsilon_{0n_0}]^\top \sim \mathcal{N}\left(0, \text{diag}_{n_1+n_0}(\sigma^2, \sigma^2, \ldots, \sigma^2)\right)$ and we again assume that the Y_{wj}'s are mutually independent. Furthermore, we denote by σ_Y^2 the variance of the outcome, Σ_X the covariance matrix of the covariates, and σ_{XY} a p-dimensional column vector consisting of the covariance between the outcome variable and each covariate.

We will again consider a superiority trial, where the $\mathcal{H}_0$ and $\mathcal{H}_1$-hypotheses are given as in equation (3.39). The ML estimates of β_0 and β_1 are normally distributed and hence their difference is normally distributed as well. By standardising the difference by the variance, we obtain a standard normal distribution under the $\mathcal{H}_0$-hypothesis. Similarly to the univariable case in equation (3.43), Zimmermann et al. [30] shows that an unbiased variance estimate is given by both expressions

$$\begin{aligned} \widehat{\mathbb{V}\text{ar}}\left(\widehat{\beta}_1 - \widehat{\beta}_0\right) &= \left(\frac{1}{n_1} + \frac{1}{n_0} + X_d^\top \left((n-2)\widehat{\Sigma}_X\right)^{-1} X_d\right)\widehat{\sigma}^2 \\ &= \left(\frac{1}{n_1} + \frac{1}{n_0} + X_d^\top \left((n-2)\widehat{\Sigma}_X\right)^{-1} X_d\right)\frac{n-2}{n-2-p}\widehat{\sigma}_Y^2(1-\widehat{R}^2), \end{aligned} \tag{3.57}$$

where $X_d := \left(\overline{X}_1^1 - \overline{X}_0^1, \ldots, \overline{X}_1^p - \overline{X}_0^p\right)^\top$. Furthermore, the estimated pooled multiple correlation coefficient between the outcome and the covariates is given as

$$\widehat{R}^2 := \frac{\widehat{\sigma}_{XY}^\top \widehat{\Sigma}_X^{-1} \widehat{\sigma}_{XY}}{\widehat{\sigma}_Y^2}, \tag{3.58}$$

where $\widehat{\sigma}_{XY}$ and $\widehat{\Sigma}_X$ are all (co)variance estimates similar to the ones in equations (3.45) and (3.42). Furthermore, $\widehat{\sigma}_Y^2$ is the standard unbiased estimator of σ_Y^2.

To determine $\widehat{\sigma}^2$ we can use the variance estimate similar to the one given in equation (3.44), but where we replace the denominator with $n-2-p$ instead of $n-3$ to obtain unbiasedness. More specifically, this estimator then follows $\frac{\sigma^2}{n-2-p}\chi^2(n-2-p)$, and thus by a similar argument as for the univariable case, it can be shown that

$$T_{\text{ANCOVA}} = \frac{\widehat{\beta}_1 - \widehat{\beta}_0 - \Delta_s}{\sqrt{\widehat{\mathbb{V}\text{ar}}\left(\widehat{\beta}_1 - \widehat{\beta}_0\right)}} \sim t(n-2-p). \tag{3.59}$$

In [30] the non-centrality parameter is shown to be

$$c(n) = \frac{\beta_1 - \beta_0 - \Delta_s}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_0} + X_d^\top \left((n-2)\Sigma_X\right)^{-1} X_d\right)\sigma_Y^2\left(1 - \frac{\sigma_{XY}^\top \Sigma_X^{-1} \sigma_{XY}}{\sigma_Y^2}\right)}}. \tag{3.60}$$

Using the non-centrality parameter, we could determine a sample size by iteratively solving

$$F^{-1}_{t,n-2-p,0}(1-\alpha/2) \leqslant F^{-1}_{t,n-2-p,c(n)}(\beta). \tag{3.61}$$

Alternatively, one could again approximate the sample size as described in the following section.

3.4.4 Approximation Formulas

We will now introduce some different approximation formulas for a one-sided test with significance level $\alpha/2$. A basic approximation formula, derived by again using that the t-distribution is approximately normal, is given by

$$n_A = \frac{(1+r)^2}{r}\frac{(z_{1-\alpha/2}+z_{1-\beta})^2\sigma^2_Y(1-R^2)}{(\beta_1-\beta_0-\Delta_s)^2}, \tag{3.62}$$

which corresponds to the multivariate version of the Frison-Pocock approximation in equation (3.54).

Again a Guenther-Schouten correction can be used when approximating the t-distribution by a normal distribution, that is

$$n_{\text{GS}} = n_A + \frac{(z_{1-\alpha/2})^2}{2}. \tag{3.63}$$

A more conservative approximation is the degrees-of-freedom adjustment given by

$$n_{\text{DF}} = n_A\frac{n_A-2}{n_A-2-p}. \tag{3.64}$$

Multiplying by this factor can be heuristically motivated by the $\frac{n_A-2}{n_A-2-p}$ factor in equation (3.57), which does not appear in the variance expression in the denominator of equation (3.60), meaning heuristically that we should adjust for it. This approximation is more conservative in the sense that $\frac{n_A-2}{n_A-2-p} > 1$ when $p \geqslant 1$, thus estimating a larger required n to obtain the desired amount of power with a given significance level. Combining the Guenther-Schouten and degrees-of-freedom corrections, we obtain an even more conservative approximation given by

$$n_{\text{GS,DF}} = n_{\text{DF}} + \frac{(z_{1-\alpha/2})^2}{2}. \tag{3.65}$$

3.5 Violations of AN(C)OVA Model Assumptions

Until now, we have assumed that the underlying data generating process is an AN(C)OVA model. Specifically, we assume e.g. that the mean value structure is linear, that the errors are homoskedastic and that the outcome is normally distributed. Based on the ANCOVA model assumptions, we have argued that we are able to provide an unbiased estimate of the ATE and,

through the t-test, to control the type I error rate as well as calculate a sample size that attains a certain level of power. Since it is necessary to assume an underlying ANCOVA model to reach these conclusions, we find it relevant to question to which degree these desirable properties hold when data is not generated from a distribution that satisfies these somewhat strict assumptions.

We will show that in an RCT, even though the ANCOVA model assumptions are violated, the ANCOVA model ATE estimator is still consistent, in the sense that $\widehat{\mathrm{ATE}} \xrightarrow{\mathbb{P}} \mathrm{ATE}$ as $n \to \infty$ [31]. Specifically, we will use the ANCOVA model ATE estimator as the second entry in $\widehat{\beta} = \left(\mathbb{D}^\top \mathbb{D}\right)^{-1} \mathbb{D}^\top \mathbb{Y}$, where the design matrix is specified as $\mathbb{D} = [1 \;\; \mathbb{W} \;\; \mathbb{X}]$.

We enforce only mild assumptions about the true data generating process within the RCT, namely that

$$\begin{aligned} X &\sim U_X, & \mathbb{E}[X] &= 0 \\ W &\sim U_W, & \mathbb{E}[W] &= \mathbb{P}(W = 1) = \pi_1 \\ Y(W) &= \mu_W(X) + \varepsilon_Y, & \mathbb{E}[\varepsilon_Y] &= 0, \end{aligned} \tag{3.66}$$

for some arbitrary distributions U_X, U_W and $\varepsilon_Y \sim U_Y$, and some arbitrary mean value structure $\mu_W(X)$. In e.g. a complete randomisation scheme, we would have $U_W = \mathrm{Bernoulli}(1/2)$. We will still assume that $W \in \{0, 1\}$ is the randomised treatment allocation and that X are baseline covariates (recorded pre-treatment) such that $W \perp\!\!\!\perp X$. Additionally, we assume that $\varepsilon_Y \perp\!\!\!\perp X$, so that all effect of X on Y is contained in $\mu_W(X)$, and that all observations are mutually independent, which is ensured in an RCT with a suitable randomisation scheme.

We note that the ANCOVA model assumptions constitute a more strict subset of these very non-exclusive assumptions. Specifically, we have not assumed any normal distribution of the error terms or linearity of the mean value structure. The assumption of $W \perp\!\!\!\perp X$ is ensured by randomisation and $\mathbb{E}[X] = 0$ can be ensured by demeaning the covariates.

In order to show consistency of the ANCOVA ATE estimator under these violations of the ANCOVA model, we first use properties of the probability limit to conclude that

$$\begin{aligned} \operatorname{plim} \widehat{\beta} &= \operatorname{plim} \left(\left(\mathbb{D}^\top \mathbb{D} \right)^{-1} \right) \operatorname{plim} \left(\mathbb{D}^\top \mathbb{Y} \right) = \operatorname{plim} \left(\mathbb{D}^\top \mathbb{D} \right)^{-1} \operatorname{plim} \left(\mathbb{D}^\top \mathbb{Y} \right) \\ &= \operatorname{plim} \left(\frac{1}{n} \mathbb{D}^\top \mathbb{D} \right)^{-1} \operatorname{plim} \left(\frac{1}{n} \mathbb{D}^\top \mathbb{Y} \right) = \mathbb{E}\left[D^\top D \right]^{-1} \mathbb{E}\left[D^\top Y \right], \end{aligned} \tag{3.67}$$

for plim denoting the probability limit as $n \to \infty$, and where the last equality follows from the law of large numbers used entry-wise [15, pp. 94–96]. Later we will derive the two factors in the last expression in equation (3.67). These are given in equations (5.55) and (5.54). From these

expressions, we get that

$$\begin{aligned}
\operatorname{plim} \widehat{\mathrm{ATE}} &= -\frac{1}{1-\pi_1}\,\mathbb{E}[Y] + \frac{1}{\pi_1(1-\pi_1)}\,\mathbb{E}[W]\,\mathbb{E}[Y(1)] \\
&= -\frac{1}{1-\pi_1}\left(\pi_1\,\mathbb{E}\left[\mu_1(X)\right] + (1-\pi_1)\,\mathbb{E}\left[\mu_0(X)\right]\right) + \frac{1}{\pi_1(1-\pi_1)}\pi_1\,\mathbb{E}\left[\mu_1(X)\right] \\
&= \frac{1}{1-\pi_1}\left(\mathbb{E}\left[\mu_1(X)\right] - \pi_1\,\mathbb{E}\left[\mu_1(X)\right]\right) - \mathbb{E}\left[\mu_0(X)\right] \\
&= \mathbb{E}\left[\mu_1(X)\right] - \mathbb{E}\left[\mu_0(X)\right] = \mathrm{ATE},
\end{aligned} \tag{3.68}$$

where we have used in the second equality that

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[WY(1) + (1-W)Y(0)] \\
&= \mathbb{E}\left[W(\mu_1(X) + \varepsilon_Y) + (1-W)(\mu_0(X) + \varepsilon_Y)\right] \\
&= \mathbb{E}[W]\,\mathbb{E}\left[\mu_1(X)\right] + \mathbb{E}[1-W]\,\mathbb{E}\left[\mu_0(X)\right] \\
&= \pi_1\,\mathbb{E}\left[\mu_1(X)\right] + (1-\pi_1)\,\mathbb{E}\left[\mu_0(X)\right],
\end{aligned} \tag{3.69}$$

thus implying that the ATE estimate is still consistent.

We can question how strict the assumption of data following a normal distribution is in regard to the t-tests allegedly following a t-distribution. For the ANOVA model t-test we can use the central limit theorem to conclude that the difference in the mean values in the numerator in (3.27) is approximately normal for sufficiently large n, even though the outcome is not normally distributed. Similarly for the ANCOVA model t-test, it has been shown that the t-test is robust to violation of the assumption of normality of the outcome [21, p. 32].

Another question regards the denominator of the t-test statistic, namely the estimated variance, under violations of the ANCOVA model. Under the assumptions of a normal linear model, including the assumption of $\Sigma = \sigma^2 I_n$, the variance of the OLS estimator is given as $\sigma^2(\mathbb{D}^\top\mathbb{D})^{-1}$. However, when these assumptions of homoskedasticity or the assumption of no correlation between errors of the observations are violated, the OLS estimator of the ATE is still unbiased (if the remaining normal linear model assumptions hold true), but inefficient due to the Gauss-Markov theorem [15, p. 107]. The variance of the OLS estimator instead takes the sandwich-form

$$\mathbb{V}\mathrm{ar}\left(\widehat{\beta}\right) = (\mathbb{D}^\top\mathbb{D})^{-1}\mathbb{D}^\top\Sigma\mathbb{D}(\mathbb{D}^\top\mathbb{D})^{-1} = \left(\frac{1}{n}\mathbb{D}^\top\mathbb{D}\right)^{-1}\frac{1}{n^2}\mathbb{D}^\top\Sigma\mathbb{D}\left(\frac{1}{n}\mathbb{D}^\top\mathbb{D}\right)^{-1}, \tag{3.70}$$

so that the asymptotic covariance matrix of the OLS estimates is given as

$$n\,\mathbb{V}\mathrm{ar}\left(\widehat{\beta}\right) = \operatorname{plim}\left(\frac{1}{n}\mathbb{D}^\top\mathbb{D}\right)^{-1}\operatorname{plim}\left(\frac{1}{n}\mathbb{D}^\top\Sigma\mathbb{D}\right)\operatorname{plim}\left(\frac{1}{n}\mathbb{D}^\top\mathbb{D}\right)^{-1}, \tag{3.71}$$

assuming that the limits exist [15, pp. 196–200]. In RCTs, we have reason to believe that the observations are independent, at least when the randomisation scheme is chosen accordingly. However, the assumption of homoskedasticity could be violated.

To account for heteroskedasticity, White [32] suggested estimating the asymptotic variance in (3.71) while taking possible heteroskedasticity into account. We will refer to the resulting covariance matrix of the OLS estimators as *heteroskedasticity-robust*, or *heteroskedasticity consistent* (HC) estimators of the OLS estimator variance. Specifically, White proved that under certain conditions, including existence of the limits, that the expression (3.71) can be consistently estimated when heteroskedasticity is present by substituting the true data covariance matrix Σ with an estimated diagonal matrix $\widehat{\Sigma} = \mathrm{diag}(\widehat{\sigma}_1^2, \widehat{\sigma}_2^2 \ldots, \widehat{\sigma}_n^2)$ with $\widehat{\sigma}_i^2$ being the squared residual of observation i. The $n \times n$ matrix Σ cannot in itself be estimated consistently, but White showed that the fixed dimension $(p+2) \times (p+2)$ "meat" matrix $\frac{1}{n}\mathbb{D}^\top \Sigma \mathbb{D}$ can indeed be estimated consistently when $n \to \infty$ using the suggested estimator $\widehat{\Sigma}$ of Σ. The limits of the remaining "bread" matrices $\left(\frac{1}{n}\mathbb{D}^\top \mathbb{D}\right)^{-1}$ follow from the law of large numbers when the covariates are well-behaved, that is, not e.g. containing a linear time trend [15, pp. 94–96]. In practice, exchanging Σ with $\widehat{\Sigma}$, the expression after the first equality in (3.70) is used [15, pp. 196–200].

To account for having a finite sample in practice when using the consistent estimator, MacKinnon and White [33] suggested different correction factors. The simplest of these is a degrees of freedom correction by the name HC_1, where we take into account the fact that the squared residuals are not unbiased estimates of the true error variance, by multiplying the meat matrix by the factor $n/\left(n-(p+2)\right)$. Other corrections, like the HC_3 correction, which is suggested to use in practice by Long and Ervin [34], inflates the residuals corresponding to observations with large leverage by multiplying the ith squared residual by $1/(1-h_i)^2$, where the leverage h_i is the i'th diagonal element of the projection matrix $P_X = X(X^\top X)^{-1}X^\top$ [33, 15, p. 200].

There seems to be no overall consensus in the literature on the necessity of these heteroskedasticity-robust estimators. Schuler et al. [35] claim that such robust estimators are in fact also robust to misspecification of the ANCOVA model, and the FDA [18] suggests to use heteroskedasticity robust estimation of the ATE variance estimates. However, according to a recent paper by Wang et al. [36], the model dependent estimators are in themselves robust to arbitrary misspecifications of the ANCOVA model in the setting of RCTs by still being consistent.

Hence, though the ANCOVA model has strict model assumptions, we see that in the setting of an RCT, violation of assumptions do not affect the consistency of the ANCOVA ATE estimator or the consistency of the model dependent standard deviation estimator, potentially making necessary steps in the form of robust estimation. Consequently, the type I error rate and sample size calculations are asymptotically robust to misspecification of the ANCOVA model.

3.6 Multiple Hypotheses Testing

So far, we have presented sample size calculations when considering power of hypothesis tests regarding a primary endpoint Y. However, clinical trials typically seek to investigate one primary

outcome as well as multiple secondary confirmatory outcomes. This means that multiple significance tests should be conducted, and when testing for multiple outcomes, a significant result is likely to occur for at least one test simply due to sampling error. For instance if m independent tests should be conducted each at significance level α, and all the corresponding $\mathcal{H}_0$-hypotheses are true, the probability of at least one type I error is $1-(1-\alpha)^m$. Thus, the required sample size should be adjusted for the purpose of correcting for this fact. Multiple ways of adjusting for the multiplicity of endpoints exist, some of which will be shortly described in the following.

One way to reduce the multiplicity problem is to clearly decide the main objective of the trial, and to make sure that the objective is not answered by multiple tests. For example when testing change in one outcome from baseline, we should only test for the change until the last measurement taken and not any intermediate measurements.

When a safety variable is part of the labeling claim, multiplicity should still be taken into account as it must be treated as a secondary confirmatory endpoint. If the trial is not specifically designed to evaluate one specific safety outcome but safety is a secondary objective, the overall safety should be assessed, and then there are often no prior hypotheses, so many plausible analyses and numerous safety findings could be of concern. This means that in the case of an adverse event (AE), p-values are of limited value to the investigator since the raised safety concern would depend on the seriousness and severity of the AE and not only the presence of an AE. Thus, a non-significant difference between the interventions would still not lead to the conclusion that the treatments are equivalent in regard to safety. In this case, multiple statistical tests would be performed to properly determine if the treatment causes potential risks, and thus an adjustment for multiplicity would be counterproductive. Therefore it is clear that in this situation there is no control over the type I error and the plausibility of any significant findings should be evaluated depending on prior knowledge of the pharmacology of the treatment [37, p. 8] [38, p. 8].

When conducting multiple tests, we wish to control the *family-wise error rate* (FWER) defined as the probability of making at least one type I error in a family of tests specified in the protocol. Two methods for controlling this error rate is introduced in the following two examples. We begin by describing the fixed testing sequence, which is appropriate when we have a specified order of the tests in the protocol, e.g. when having primary and secondary confirmatory outcomes.

Example 3.6.1. Fixed testing sequence
In trials where the confirmatory endpoints can be ordered in regard to for example clinical meaningfulness, a procedure called *fixed testing sequence* can be used. The endpoints are tested in a predetermined sequence specified in the protocol, meaning that the hypotheses are ordered as $\mathcal{H}_0^{(1)}, \mathcal{H}_0^{(2)}, \ldots, \mathcal{H}_0^{(m)}$. Then each $\mathcal{H}_0^{(i)}$, $i = 1, 2, \ldots, k \leqslant m$ is tested one at a time, starting from $i = 1$, at the same significance level α, until the first non-significant result k is obtained. That is, the hypotheses are tested in a prespecified order until $\mathcal{H}_0^{(k)}$ is not rejected. This procedure is also referred to as a closed test procedure [39]. In proposition B.1.1 in appendix B.1, we show that the FWER is bounded by α for this procedure. ◀

When there is no natural ordering of the hypotheses $\mathcal{H}_0^1, \mathcal{H}_0^2, \ldots, \mathcal{H}_0^m$, we need an alternative to the fixed testing sequence procedure, which we present in the next example.

Example 3.6.2. Bonferroni corrections
The simple Bonferroni method is to reject $\mathcal{H}_0^i$ if the associated p-value fulfills $p_i < \alpha_i$ with $\sum_{i=1}^m \alpha_i = \alpha$. Often $\alpha_i = \alpha/m$, which yields a procedure called the *unweighted simple Bonferroni method*.

A mix of the Bonferroni procedure and the fixed testing sequence procedure is the *Holm's sequentially rejective Bonferroni method*. The testing sequence is sorted according to the lowest to highest p-value which we denote by $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$. According to this procedure, the hypotheses are tested sequentially, and $\mathcal{H}_0^{(i)}$ is rejected if $p_{(i)} < \alpha/(m-i+1)$. If $\mathcal{H}_0^{(i)}$ is not rejected, the procedure stops, and otherwise proceed to test $\mathcal{H}_0^{(i+1)}$.

Both procedures have the FWER bounded by α, as we show in propositions B.2.1 and B.2.2 in appendix B.2. ◀

As described by equation (3.4), decreasing the probability of type I error decreases the power when using the same n. This implies that when using the unweighted simple Bonferroni method, the power suffers increasingly with m, being the price of controlling the FWER while being able to test all the hypotheses. For the Bonferroni-Holm procedure, we control the FWER while increasing the power compared to the simple approach, but we run the risk of not being able to test all hypotheses. For the fixed testing sequence procedure, we again control the FWER, and since all tests are performed on a significance level of α, the power does not suffer. However, the Bonferroni-Holm procedure makes it more likely to be able to test all hypotheses, since we are allowed to ascendingly arrange the p-values, but each test is more easily non-rejected due to the smaller significance level [40, 41].

For the fixed sequence procedure and the simple Bonferroni correction, we have prefixed significance levels at which each hypothesis is tested. On the other hand, the Bonferroni-Holm procedure is dependent on data, since the level applied to a specific hypothesis is not known beforehand. Thus, it can be more challenging to determine the sample size for the Bonferroni-Holm procedure compared to the fixed sequence procedure and the simple Bonferroni procedure since it is not known which significance level should be applied to the specific $\mathcal{H}_0$-hypothesis in the planning phase of a trial. Vickerstaff et al. [42] suggest to use a simulation based approach when the levels are data driven.

We can derive sample size calculation methods applicable for the prefixed level procedures. To be able to do so, we need to extend the definition of power for multiple hypotheses testing, which can be done in several ways. The probability of rejecting all false $\mathcal{H}_0$-hypotheses is called the *conjunctive power*, whereas the probability of rejecting at least one false $\mathcal{H}_0$-hypothesis is called the *disjunctive power*.

Example 3.6.3. Sample size calculation for prefixed level procedures
We let the prefixed levels be given by α_i for $i = 1, 2, \ldots, m$, such that the testing procedure ensures that $\mathrm{FWER} \leqslant \alpha$ for a prespecified significance level α. Since the power is the probability of correctly rejecting a false $\mathcal{H}_0$-hypothesis, we will assume that all $\mathcal{H}_0$-hypotheses are false in order to calculate the power.

If the aim is to ensure a conjunctive power of at least $1-\beta$, we can calculate the sample size based on the significance level α_i and power $1-\beta/m$ for all $i = 1, 2, \ldots, m$. Then taking the maximum of all of these sample sizes, we would obtain a conjunctive power of at least $1-\beta$. This follows since under the specified alternative hypotheses, we have

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i=1}^{m}\{p_i < \alpha_i\}\right) &= 1 - \mathbb{P}\left(\bigcup_{i=1}^{m}\{p_i \geqslant \alpha_i\}\right) \geqslant 1 - \sum_{i=1}^{m}\mathbb{P}(p_i \geqslant \alpha_i) \\ &= 1 - m \cdot \frac{\beta}{m} = 1 - \beta.\end{aligned} \tag{3.72}$$

In the inequality we use the Bonferroni inequality, which generally states that for a countable set of events $A_1, A_2, A_3, \ldots$,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leqslant \sum_{i=1}^{\infty}\mathbb{P}(A_i). \tag{3.73}$$

In the second equality we use that under the alternative hypothesis the probability of conducting a type II error is β/m.

To obtain a disjunctive power of at least $1-\beta$, we can calculate the sample size based on the significance level α_i and power $1-\beta$ for all $i = 1, 2, \ldots, m$. Again we should then take the maximum of all of these sample sizes. This follows since under the alternative hypotheses, we then have

$$\mathbb{P}\left(\bigcup_{i=1}^{m}\{p_i < \alpha_i\}\right) \geqslant \mathbb{P}\left(p_k < \alpha_k\right) = 1 - \beta, \tag{3.74}$$

for any $k \in \{1, 2, \ldots, m\}$ [21, pp. 133–135]. ◀

4 | Synthetic Control Arms

In this chapter, we will describe an approach to using historical data, both in the setting where the current trial includes only a treatment arm and the setting where a control arm is present as well. Namely, we will describe the approach of using synthetic control arms (SCAs). When using historical controls in single-arm trials as an external comparator, it enables estimation of the ATE. In the case of a current control group being populated with historical control group patients, we get a larger n, so the variance of the treatment estimate decreases, and thus the power increases.

Throughout this thesis we assume that we have access to a historical data set of n' independent and identically distributed observations $(\mathbb{X}', \mathbb{Y}') \in \mathcal{X}^{n'} \times \mathbb{R}^{n'}$, which are also independent from the observations in the current trial data set. Recall that the covariates X in the current RCT fulfills $X \in \mathcal{X}$, meaning that the same covariates are recorded in the historical dataset as in the current RCT. However, the covariate distribution in the historical data is potentially different than that in the current RCT. The "best-case" scenario is $p(X', Y') = p\left(X, Y(0)\right)$, where p denotes the joint distribution. In this case, where the joint distribution of a subject's covariates and (potential control) outcome is the same for the historical and current RCT patients, the historical population gives maximal information on the RCT control arm. This best case scenario rarely presents itself in practice, but methods presented in this and the next chapter can be utilised to increase power, ideally even though the distributions deviate from each other.

When using the SCA approach, we seek to populate the data set with similar patients as those in the treatment group with control group patients of historical clinical trials, either in the situation where a current control group is present or in the context of a single-arm trial. However, in both cases the estimated treatment effect is susceptible to bias if we are not cautious about which historical control group patients we choose to include, which has risk of inflating the type I error [43]. That is, specific to this approach, where we seek to use external controls directly in the analysis, the historical data should be comparable to the current data. Specifically, Pocock [44] suggested that in the context of using historical data with a current RCT, the historical control group should meet the following six criteria:

" 1. Such a group must have received a precisely defined standard treatment which must be the same as the treatment for the randomized controls.

2. The group must have been part of a recent clinical study which contained the same requirements for patient eligibility.

3. The methods of treatment evaluation must be the same.

4. The distributions of important patient characteristics in the group should be comparable with those in the new trial.

5. *The previous study must have been performed in the same organization with largely the same clinical investigators.*
6. *There must be no other indications leading one to expect differing results between the randomized and historical controls."*

In general, these conditions are very strict in the sense that they narrow the data which is useful for historical controls down to RCTs which are highly similar to the current RCT, ensuring that bias does not occur when comparing the outcome of treatment and control group patients.

As we will elaborate in the following, there are methods that seek to relax some of the assumptions. One class of approaches for doing so is based on using the *propensity score*, presented in the next section, which is used to match the historical control group patients to the treatment population in order to mimic this group in regard to known confounders. Using matching based on the propensity score, patients from control groups of previously conducted clinical trials can be matched to current treatment group patients based on a propensity score. The matched patients originating from historical control groups then constitute an SCA, which can be used as an external comparator arm for the treatment arm.

Ideally, matching by the propensity score will ensure unbiasedness of the treatment estimate. Criterion 2 can be fulfilled by matching with only the patients from historical control groups that are eligible in regard to the current RCT. Furthermore, as we will show, the criteria relaxed by using propensity score matching for constructing a synthetic control arm are criterion 4 (by matching patients that are similar to those in the treatment group), 5 and 6 (by modelling the bias between the current and historical treatment groups, when a current control group is present).

4.1 Propensity Score Matching

Propensity score matching is a subcategory of general methods with an overall goal of controlling for differences in observed confounders between a treatment and control group. The control for confounders is to ensure that the observed differences in outcome between a treatment group and some historical controls can be attributed to the treatment itself. In other words, the method seeks to remove (or reduce) the bias induced by comparing two groups that are dissimilar in regard to observed baseline confounders in order to enable causal inference [43]. The method was originally intended for non-randomised comparative studies such as the situation where all patients are exposed to treatment and we wish to compare their outcome with historical controls. However, the method has been developed such that it can be used in the context where some concurrent randomised control patients are present as well [43, 45].

The overall idea of propensity score methods is to account for differences between two groups, namely the treatment and control groups, making them comparable. Formally, this could be achieved by a *balancing score*, which Rosenbaum and Rubin [46] defines as a function b satisfying

$$X \perp\!\!\!\perp W \mid b(X), \tag{4.1}$$

where X denotes (confounding) covariates. That is, the balancing score is a function such that the conditional distribution of $X \mid b(X)$ is the same for patients in the treatment- and control groups, respectively. In other words, b is a function satisfying that there is no difference in the distribution of (confounding) covariates X between the treatment- and control groups when $b(X)$ is given.

A trivial choice of b is the identity $b(x) = x$, which makes statement (4.1) true. One way to use this in practice would be matching patients from the treatment group with patients in the historical control groups based on x, being all confounding covariates. However, this quickly becomes infeasible as the number of confounders grows, due to the curse of dimensionality. This is why a *scalar balancing score*, $b\colon \mathcal{X} \to \mathbb{R}$, is preferred. Rosenbaum and Rubin [46] showed that the propensity score, defined by

$$e(x) := \mathbb{P}\left(W = 1 \mid X = x\right), \tag{4.2}$$

where x are baseline confounders, is in fact a balancing score. Intuitively, this is a balancing score since matching a historical control based on the probability of that control being in the treatment group, each pair of matched patients could, based on the covariates, just as well have their treatment assignments swapped, entailing that the treatment allocation does not depend on X. In practice, the propensity score is estimated using a propensity score model. This model is fitted using all patients, and is estimated for all patients. That is, the propensity score is estimated for patients both in the treatment group as well as in the historical controls and, if present, current controls. The model for estimating the propensity score can be chosen freely among e.g. traditional models such as logistic or probit regression, or models such as classification trees and forests [7].

Austin [7] describes several methods for using the propensity score, including matching, stratification, inverse probability of treatment weighting and covariate adjustment. Among these methods, matching has been shown to achieve best results in reducing differences in confounders between the treatment and control groups when we have access to a rich amount of high-quality historical data [43, 7]. Therefore we will focus on this method. In the following, we will describe methods for matching on the propensity score both in the case of single-arm trials, where no current controls are present, and in the case of two-arm trials where a current control group is present.

In case of both a single-arm and multi-arm trial, we wish to estimate the ATE in the population in the current trial. In the case of a single arm trial, the population consists of only patients in the treatment group, meaning that ATT among all patients and ATE among the current trial patients are the same. For the two-arm setting, we see from the condition in (2.5) that ATT and ATE coincides in a randomised setting, that is, within the current trial population. Furthermore, the ATT among patients in the current trial and all patients coincide. That is, in both cases, our estimand of interest is the ATT among all patients.

We will now introduce an important assumption that is needed in order to estimate the ATT. We

say that the treatment assignment is *strongly ignorable* given X, if

$$\big(Y(0), Y(1)\big) \perp\!\!\!\perp W \mid X \qquad \text{and} \qquad 0 < \mathbb{P}\big(W = 1 \mid X\big) < 1. \tag{4.3}$$

If this is the case, then treatment assignment is also strongly ignorable given any balancing score, in particular the propensity score. That is,

$$\big(Y(0), Y(1)\big) \perp\!\!\!\perp W \mid e(X) \qquad \text{and} \qquad 0 < \mathbb{P}\big(W = 1 \,\big|\, e(X)\big) < 1. \tag{4.4}$$

This important property ensures that at any given value of the propensity score, the potential outcomes of a patient in the treatment group is directly comparable to the potential outcomes of a patient in the control group with the same propensity score. This ensures that the difference between outcomes of such patients can be ascribed to the treatment assignment. Thus, matching each treatment group patient to one or more control group patients on their propensity score entails being able to perform causal inference between the outcome and treatment assignment.

We can think of the matching based on the propensity score as controlling for the tendency of e.g. older patients to be in the historical controls more frequently than in our treatment group. It would presumably affect the results if we e.g. used a random subset of the controls instead of matching using the propensity score. In this case, matching on the propensity score, we control for the confounding effect of age. The procedure is called pseudo-randomisation because we match patients in the treatment group with similar patients in the control groups. In that way, the method seeks to ensure that the included patients have an equal chance of being in the treatment and control group, as if they were randomly assigned to treatment or control in regard to the confounders included in the propensity score.

Formally, we will use the assumption of strong ignorability to obtain an unbiased estimate of the ATT. We will later see that the unbiased estimate is obtained using the relation

$$\begin{aligned}
\text{ATT} &= \mathbb{E}\big[Y(1) \,\big|\, W = 1\big] - \mathbb{E}\big[Y(0) \,\big|\, W = 1\big] \\
&= \mathbb{E}_{e(X)\,|\,W=1}\Big[\mathbb{E}\big[Y(1) \,\big|\, e(X), W = 1\big] - \mathbb{E}\big[Y(0) \,\big|\, e(X), W = 1\big]\Big] \\
&= \mathbb{E}_{e(X)\,|\,W=1}\Big[\mathbb{E}\big[Y(1) \,\big|\, e(X), W = 1\big] - \mathbb{E}\big[Y(0) \,\big|\, e(X), W = 0\big]\Big],
\end{aligned} \tag{4.5}$$

where the assumption of strong ignorability is used in the last equality [46, 6].

In practice, we will need to justify the assumption of strong ignorability given X, since we then have fulfillment of the necessary assumption of strong igonrability given $e(X)$. This will then ensure that the above relation can be used to find an estimator of the ATT. However, the assumption of strong ignorability given X is indeed a strong assumption. By equation (2.3) the first part of the assumption implies that the treatment assignment is unconfounded with $\big(Y(0), Y(1)\big)$ given X. This is referred to as *the assumption of unconfoundedness* [5]. This entails that all confounders of the treatment effect are contained in X. This is the reason for also referring to the assumption as *the assumption of no unobserved confounders* [46]. The assumption of no unobserved confounders is rarely reasonable in practice, and this will also disturb the causal inference.

The second part of the assumption is referred to as *the assumption of overlap*, since it entails that patients have a positive probability of belonging to both the treatment and control group across the whole sample space of the confounders. This is to ensure that the propensity score can be used e.g. to find relevant matches in the historical data for every patient in the treatment group.

We note, however, that using the propensity score in general requires the propensity score model to be adequately specified, since $e(X)$ is not known in practice and hence need to be estimated. If the estimated propensity score does not resemble the true propensity score, the relation in equation (4.4) cannot be expected to hold when we use an estimate of $e(X)$. Another problem that arises in practice when using for example the logistic model is that it is not possible to test the assumption of overlap directly from the model. This is due to the model never estimating probabilities of exactly 0 and 1. In section 4.2, we return to some methods for assessing the propensity score model.

4.1.1 Estimating the ATT

The method of propensity score matching relies on matching (preferably) each of the patients in the treatment group with one or more patients in a pooled group of possible controls. In the context of a single-arm trial, this group consists of some historical controls, and in the context of a two-arm trial, it refers to the pooled group of current controls and historical controls. Even though our aim will be to construct an actual control group from this pool of possible control group patients, we will refer to the pooled group of possible controls simply as the control group or the controls.

Matching can be carried out in several ways, here exemplified in the case of *one-to-one matching*, where each treated patient is matched to a single control group patient based on the propensity score. For *greedy matching*, patients in the treatment group are selected randomly in turn and matched with the patient in the control group with the smallest absolute difference according to the propensity score. This is carried out until every patient in the treatment group is matched to a unique control group patient. It can also be carried out by *optimal matching*, where the sum of absolute deviations across all pairs are minimised. Comparison between the two methods suggests that optimal matching does not perform better than greedy matching in regard to balancing covariate values between treatment and control group patients [7, 47].

These procedures do not guarantee that treatment group patients are in fact matched with highly similar control group patients, since there is no guarantee of patients with highly similar propensity scores exist in the control group. This is why a rich amount of historical data is needed for this method to work well. However, we can use a *caliper*, which is a pre-specified range of the propensity score that patients can be matched within. If no patients in the control group has a propensity score within the caliper of a treatment group patient's propensity score, we will then need to discard this treatment patient from the analysis. For continuous outcomes, a caliper width of 0.2 times the standard deviation of the logit of the propensity score have been shown to work well [48].

The methods just described use matching without replacement. Matching with replacement is also possible, but the variance estimate of the treatment effect then needs to be adjusted according to the resulting correlation structure since more of the same information is used [7].

When matching has been carried out, we can use the relation in equation (4.5) to obtain an unbiased estimate of the ATT as

$$\widehat{\mathrm{ATT}} = \frac{1}{n_1} \sum_{i:w_i=1} \left(y_i(1) - \frac{1}{M_i} \sum_{j=1}^{M_i} \widehat{y}_{i,j}(0) \right), \tag{4.6}$$

where M_i is the number of patients from control groups matched (based on the propensity score) to patient i in the treatment group, having outcomes $\widehat{y}_{i,j}(0)$ [5, 43]. That is, we seek to estimate the counterfactual (unobserved) outcome $y_i(0)$ of the treatment group patients with the observed outcome of patients in the control group, which are similar according to (preferably all) confounding covariates.

The fact that the entity in equation (4.6) reasonably estimates the ATT can be seen from equation (4.5). Here, the outer expectation with regard to the distribution of $e(X) \mid W = 1$ is estimated through the sample mean of outcomes for patients having propensity scores distributed as in the treatment group. The entities $\mathbb{E}\left[Y(1) \mid e(X), W = 1\right]$ and $\mathbb{E}\left[Y(0) \mid e(X), W = 0\right]$ are then estimated through the observed $y_i(1)$ and $\frac{1}{M_i} \sum_{j=1}^{M_i} \widehat{y}_{i,j}(0)$, which are all outcomes for patients having similar propensity scores $e(x_i)$ since they are matched based on this criterion. Estimating the expectations using sample means, we obtain unbiasedness of the estimate if we were able to use the true propensity scores.

A perhaps more intuitive way to see that the ATT is estimated from equation (4.6), comes from the fact that we discard from the estimate the observations in the control group that are not similar to the patients in the treatment group [45]. In a similar manner, control group patients could be matched to a treatment group patient (and discarding the remaining treatment group patients from the analysis), thereby obtaining an estimate of ATC.

The estimator in equation (4.6) implicitly assumes that there exists at least one match for each patient in the treatment group. If such a match does not exist, e.g. in the situation where a narrow caliper is used, we are forced to discard treated patients without a match, based on their propensity score $e(X)$. Thus, we do not get an appropriate estimate of the ATT since the propensity scores cannot be assumed to be realisations of the distribution of $e(X) \mid W = 1$. This is again a reason why a rich amount of data is needed for the method to work well. In such a situation, we are able to choose M_i large while maintaining similarity between the propensity scores between treated patients and their matches, and thus the variance of the estimator decreases.

Example 4.1.1. Fixed number of matches
In the case of choosing a fixed number $M_i = M$ of matches for each treatment group patient $i = 1, \ldots, n_1$, we get that the estimate corresponds to fitting a simple ANOVA model to equation

(2.17), where control group patients have been added to the data set. This is seen since we then have

$$\begin{aligned}\widehat{\text{ATT}} &= \frac{1}{n_1}\sum_{i:w_i=1}\left(y_i(1) - \frac{1}{M}\sum_{j=1}^{M}\widehat{y}_{i,j}(0)\right) \\ &= \frac{1}{n_1}\sum_{i:w_i=1} y_i(1) - \frac{1}{n_1 M}\sum_{i:w_i=1}\sum_{j=1}^{M}\widehat{y}_{i,j}(0) \\ &= \overline{y(1)} - \overline{\widehat{y}(0)} = \widehat{\beta}_W,\end{aligned} \tag{4.7}$$

which is the standard ANOVA treatment estimate of the pooled data as seen in appendix A.1. Choosing $M_i = 1$ for all $i = 1, \ldots, n_1$, we retrieve an estimate of the ATT based on one-to-one matching. From equation (4.7), we see that in this situation, the one-to-one matching ANOVA estimate reduces to

$$\widehat{\text{ATT}} = \frac{1}{n_1}\sum_{i:w_i=1}\left(y_i(1) - \widehat{y}_i(0)\right). \tag{4.8}$$

As shown earlier, benefits in terms of reduced variance of the ATT estimator can be achieved by adjusting for chance imbalances with an ANCOVA model instead of an ANOVA model. In the context of matching on propensity scores, the covariate imbalances between the treatment- and matched control groups are due to chance from the property in equation (4.1), and we would hence not expect the treatment estimate to change when using the ANCOVA estimator instead of the ANOVA estimator. ◀

When estimating the estimator in equation (4.6), there is no consensus in the literature as to how the synthetic control arm patients should be regarded. Some argue that the treated and untreated patients should be considered as $n_1 + \sum_{i=1}^{n_1} M_i$ independent observations, causing the usual variance estimate of the ML estimator to be appropriate. Others argue that the observations are dependent since the control group patients are chosen through the propensity score. In this latter case, the variance estimate and significance test of the treatment estimate could consider data as a paired sample of n_1 paired observations $y_i(1)$ and $\frac{1}{M_i}\sum_{j=1}^{M_i}\widehat{y}_{i,j}(0)$, running an AN(C)OVA model t-test on the observations $y_i(1) - \frac{1}{M_i}\sum_{j=1}^{M_i}\widehat{y}_{i,j}(0)$ [7, 5]. This would increase the variance estimate compared to regarding the observations as independent, since we would decrease n by regarding matched observations as if they were the same individual. Another approach to modeling the covariance between observations is described in example 4.1.2.

Example 4.1.2. Cluster-robust estimation of the ATE variance
Another approach to handling correlation between matched observations is to model the covariance matrix Σ of ε in the ANOVA model (2.17) or the ANCOVA model (2.20) accordingly while still using the ATE estimate $\widehat{\beta}_W$ [49].

As we have seen, under violations of the assumptions of the normal linear model that the error terms are homoskedastic and have no autocorrelation, the variance of the OLS estimator takes the sandwich form in (3.70). It is possible to estimate the $(p+2)\times(p+2)$ matrix $A \coloneqq \mathbb{D}^\top \Sigma \mathbb{D}$ assuming not only heteroskedasticity, but also a cluster-structure of the observations, where non-zero covariances between matched patients are modelled. We group matched patients in the clusters $g = 1, 2, \ldots, G$ and denoting by $\mathbb{Y}_g$ the response column vector of patients in cluster g, $\mathbb{D}_g$ the design matrix with a row for each patient in cluster g and ε_g the error column vector for each patient in cluster g. We denote the corresponding reordered entities as $\widetilde{\mathbb{D}} = \left[\mathbb{D}_1^\top \ \mathbb{D}_2^\top \ \ldots \ \mathbb{D}_G^\top\right]^\top$, $\widetilde{\mathbb{Y}} = \left(\mathbb{Y}_1^\top, \mathbb{Y}_2^\top, \ldots, \mathbb{Y}_G^\top\right)^\top$ and $\widetilde{\varepsilon} = \left(\varepsilon_1^\top, \varepsilon_2^\top, \ldots, \varepsilon_G^\top\right)^\top$. Hence, we can represent the model as

$$\widetilde{\mathbb{Y}} = \widetilde{\mathbb{D}}\beta + \widetilde{\varepsilon}, \tag{4.9}$$

and then, according to Cameron et al. [50], under suitable conditions, we can consistently estimate A by

$$\widehat{A} = \widetilde{\mathbb{D}}^\top \widehat{\widetilde{\Sigma}} \widetilde{\mathbb{D}} = \widetilde{\mathbb{D}}^\top \begin{bmatrix} \widehat{\varepsilon}_1\widehat{\varepsilon}_1^\top & 0 & \cdots & 0 \\ 0 & \widehat{\varepsilon}_2\widehat{\varepsilon}_2^\top & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \widehat{\varepsilon}_G\widehat{\varepsilon}_G^\top \end{bmatrix} \widetilde{\mathbb{D}} = \sum_{g=1}^{G} \mathbb{D}_g^\top \widehat{\varepsilon}_g \widehat{\varepsilon}_g^\top \mathbb{D}_g. \tag{4.10}$$

When using this estimator of A in equation (3.70), we will refer to the method as *cluster-robust estimation* of the covariance matrix. Heuristically, consistency of $\widehat{A}$ is due to $\widehat{\widetilde{\Sigma}}$ becoming relatively more sparse as $G \to \infty$ [50]. To take into account that in practice, the sample size is finite, the cluster-robust estimator can be corrected by the correction factors suggested for the HC estimators introduced in section 3.5. ◀

We can use the methods described in section 3.3, 3.4.1 and 3.4.3 to perform sample size calculations for PSM since estimating the treatment effect just reduces to using ANOVA or ANCOVA and potentially using robust estimation of the standard deviation.

4.1.2 Propensity Score Matching in Clinical Trials

In the single-arm case, the method just described constructs a synthetic control arm with data from historical control groups as an external comparator in order to estimate the treatment effect in a current clinical trial without a control arm. In the two-arm setting, we can use propensity score matching to improve the power in estimating the treatment effect in a current RCT, where some control group patients are already present. In this section, we will describe an extension of a procedure, which ensures the use of all current control group patients that have been part of the randomisation in the current clinical trial.

So far, the described matching procedure have not taken into account the potential outcome differences between different sources of historical control groups due to unobserved confounders.

Stuart and Rubin [45] proposed a procedure for one-to-one propensity score matching in the setting of observational studies where data from multiple external control groups are available. The procedure involves matching the patients in the treatment group to historical controls on observed confounders, as well as estimating the outcome bias between control groups induced by unobserved confounders. It is then possible to use this estimated bias between control groups with the aim of removing the bias in the estimated treatment effect by implicitly taking these unobserved confounders into account.

The procedure was adapted to the setting of confirmatory clinical trials by Lim et al. [43], ensuring that all control arm patients from the current RCT are used in the treatment effect estimation in equation (4.6). We describe the adapted procedure in further details in the following, where we denote by AT, CC and HC the active treatment arm, the current control arm and the pool of possible historical controls, respectively. Furthermore, we assume that $|\mathrm{AT}| > |\mathrm{CC}|$ and $|\mathrm{HC}| > |\mathrm{AT}| - |\mathrm{CC}|$. The procedure can be described as follows:

1. Estimate a model for the propensity scores of belonging to AT, using all patients from AT, CC and HC. That is, the CC and HC groups are pooled to one large control group. These propensity scores are used in all of the following steps.

2. Construct the group of matched patients $\mathrm{CC}:\mathrm{AT}$ by the following procedure:

   - Calculate all pairwise differences between propensity scores of patients in CC and AT.
   - Match the pair with the smallest difference in propensity scores. In case of ties, choose one pair randomly among the tied pairs with lowest differences in propensity scores.
   - Iterate the above step, but consider only unmatched pairs in each iteration. Iterate until every patient in CC is matched with a unique patient in AT.
   - Name the group of matched patients from AT and CC by $\mathrm{CC}:\mathrm{AT}$.

3. Construct the group of patients $\mathrm{CC}:\mathrm{HC}$ consisting of matched patients in CC and HC, where all patients in CC is matched with a unique patient in HC. Use the procedure described in step 2.

4. Construct the group of matched patients $\mathrm{AT}:\mathrm{HC}$, considering only the subset of patients in AT which are not matched with patients in CC in step 2. All patients in this subset of AT should be matched with a unique patient in HC. Use the procedure described in step 2.

5. Estimate the bias in control group outcome in HC by using the observations in CC as reference. Specifically, assume that the outcome is normally distributed and fit a normal linear model on the $\mathrm{CC}:\mathrm{HC}$ data, specified as

$$\mathbb{Y}(0) = \alpha + \delta \mathbb{I}_{\mathrm{HC}} + \mathbb{X}\gamma + \varepsilon, \tag{4.11}$$

where $\mathbb{I}_{\text{HC}}$ is the column vector of indicators of whether the patient is in HC, $\mathbb{X}$ is the design matrix with p columns being the covariates used for the model of the propensity score, and $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$. Denote the estimated bias as $\widehat{\delta}$. This is an estimate of the difference in the outcomes between HC and CC patients. The intention is to estimate the outcome bias among HC patients induced by factors not included in $\mathbb{X}$, that is, unobserved confounders.

The reason for only using patients from HC that are matched to patients in CC is to not impose the linearity assumption of the effect of X on Y to the complete space of X in HC, but instead to impose it to only the subset of HC where X is similar between CC and HC.

Another implicit assumption made in this step is that the group effect δ is constant across all values of X, that is, there is no interaction effects. In practice we could choose to specify a model that accounts for this, but simulation studies indicate that the method is not sensitive to violation of this assumption [45]. Heuristically, this seems reasonable since the covariate distributions among the CC and HC patients in the CC : HC group are similar.

6. In order to take the uncertainty of estimating the bias δ in step 5 into account, we now wish to sample estimates of the bias term a number of times using the estimated linear model. Denoting the design matrix for this model as $\mathbb{D} = [1 \;\; \mathbb{I}_{\text{HC}} \;\; \mathbb{X}]$, we have that

$$\widehat{\delta} \sim \mathcal{N}\left(\delta, \sigma^2 \left[(\mathbb{D}^\top \mathbb{D})^{-1}\right]_{2,2}\right) \tag{4.12}$$

$$\widehat{\sigma}^2 \sim \frac{\sigma^2}{n_c - (p+2)} \chi^2\left(n_c - (p+2)\right). \tag{4.13}$$

from standard results of the normal linear model (4.11).

Instead of sampling an estimate of δ, we can take a Bayesian approach and regard δ and σ^2 as being stochastic, and sample them instead. Specifically, we wish to sample around $\widehat{\delta}$, using a realisation of the variance $\sigma^2 \left[(\mathbb{D}^\top \mathbb{D})^{-1}\right]_{2,2}$. If we denote by n_c the number of patients in CC : HC and by Scale-inv-χ^2 the scaled inverse χ^2 distribution, we get from equation (4.13) that

$$\begin{aligned} \left(\frac{\sigma^2}{\left(n_c - (p+2)\right)\widehat{\sigma}^2}\right)^{-1} &\sim \chi^2\left(n_c - (p+2)\right) \\ \Rightarrow \; \sigma^2 &\sim \text{Scale-inv-}\chi^2\left(n_c - (p+2), \widehat{\sigma}^2\right). \end{aligned} \tag{4.14}$$

Having sampled a value s^2 from Scale-inv-$\chi^2\left(n_c - (p+2), \widehat{\sigma}^2\right)$, we can sample an estimated bias d from the conditional distribution

$$\delta \mid s^2 \sim \mathcal{N}\left(\widehat{\delta}, s^2 \left[(\mathbb{D}^\top \mathbb{D})^{-1}\right]_{2,2}\right). \tag{4.15}$$

7. We will now estimate the ATT, using the matched groups $\mathrm{AT}:\mathrm{CC}$ and $\mathrm{AT}:\mathrm{HC}$, correcting the outcomes of the historical control group patients with the estimated bias from step 6. We can estimate it from e.g. the ANOVA estimate in equation (4.8) by using that for each treated patient i in AT, we have

$$\widehat{y}_i(0) = \begin{cases} y_i^{\mathrm{CC}}(0) & \text{if the matched patient is in CC} \\ y_i^{\mathrm{HC}}(0) - d & \text{if the matched patient is in HC} \end{cases}, \tag{4.16}$$

where $y_i^{\mathrm{CC}}(0)$ denotes the outcome of the patient in CC matched with patient i in $\mathrm{AT}:\mathrm{CC}$, and equivalently for $y_i^{\mathrm{HC}}(0)$. That is, we correct for the bias induced by using a HC control group patient rather than a control group patient from the CC control group. We then estimate the ATT by an appropriate regression method, from which we also get an estimate $\widehat{u}$ of the variance u of $\widehat{\mathrm{ATT}}$. Here, a cluster-robust estimation of $\mathbb{V}\mathrm{ar}(\widehat{\mathrm{ATT}})$ can be used.

8. By iterating steps 6-7 B times, we provide a data-driven way of taking the uncertainty of estimating the bias into account when estimating ATT and the variance of the estimator. Specifically, we let $\widehat{\mathrm{ATT}}_b$ be the bth treatment estimate, and we then estimate ATT by the empirical mean

$$\widehat{\mathrm{ATT}} = \frac{1}{B}\sum_{b=1}^{B} \widehat{\mathrm{ATT}}_b. \tag{4.17}$$

Stuart and Rubin [45] suggest to estimate the variance of $\widehat{\mathrm{ATT}} - \mathrm{ATT}$ by

$$\widehat{\mathbb{V}\mathrm{ar}}\left(\widehat{\mathrm{ATT}} - \mathrm{ATT}\right) = \frac{1}{B}\sum_{b=1}^{B} \widehat{u}_b + \left(1 + \frac{1}{B}\right) \cdot \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\mathrm{ATT}} - \widehat{\mathrm{ATT}}_b\right)^2, \tag{4.18}$$

where $\widehat{u}_b$ is the estimated within-sample variance estimate of $\widehat{\mathrm{ATT}}_b - \mathrm{ATT}$, and the second factor in the second term is the between-sample variance. Using an ANCOVA model to estimate each $\widehat{\mathrm{ATT}}_b$ instead of an ANOVA model can potentially decrease both the within-sample and between-sample variance. Assuming that $\widehat{\mathrm{ATT}}_b$, estimated from a normal linear model, is normally distributed, $\widehat{\mathrm{ATT}}$ is also normally distributed.

In the original procedure proposed by Stuart and Rubin [45], the $\mathrm{CC}:\mathrm{AT}$ group is constructed in a more general way by an *extended caliper* method, where only patients in CC with a propensity score within a pre-specified margin of the propensity scores of patients in AT are considered. The rest of the patients in AT are then matched with patients in HC. In this original formulation, the size of the extended caliper reflects the preference for matching patients from AT with patients in CC rather than HC. That is, for small extended calipers, patients in CC have the risk of being discarded, which is arguably a bad choice when CC consists of patients randomised to the control group rather than being a non-randomised control cohort. In this regard, the modification made by Lim et al. [43] consists of choosing this extended caliper to be ∞.

An assumption for the original procedure to work is that the treatment assignment mechanism into AT and CC is strongly ignorable given X. In practice, strong ignorability with regard to X means that AT and CC are comparable when we control for X [45]; that is, no confounding covariates are present apart from those in X. Heuristically, this means that the bias-correction in step 7 of the procedure ensures comparability between AT and the part of HC we match to AT, since we match on the confounders in X. Furthermore, the bias-correction ensures comparability to CC, which is again comparable to AT. The reason for the bias-correction is then that strong ignorability is not assumed for treatment assignment into the group of patients in $\mathrm{AT}:\mathrm{HC}$. In practice, this is not fulfilled since HC could be obtained from different geographical regions or different clinics. We notice that in the setup of CC coming from a randomised trial, the assumption of strong ignorability between AT and CC is fulfilled due to randomisation [46].

We note that even though the method described here seeks to take potential bias induced by unobserved confounders into account, no existing research seems to have provided any analytical guarantee that the type I error rate is controlled. We suspect that type I error rate inflation could occur due to bias in the ATE estimate induced by misspecification of the propensity score model as well as the quality and distribution of historical data.

4.2 Balance Diagnostics and Variable Selection

Before using propensity score matching to estimate the ATT, the quality of the obtained matches should be assessed by using *balance diagnostics*. These are tools for evaluating whether the condition in equation (4.1) is fulfilled for the estimated propensity score. That is, we wish to examine whether the baseline confounders X are distributed equally among the group of treated patients and the group of matched control group patients. We know that the condition holds for the true propensity score $e(X)$. Thus, non-fulfillment of the condition suggests that good matches for all treatment group patients do not exist in the historical control group, or the model used for the propensity score is wrongly specified. In the latter case, the propensity score model needs to be revised, e.g. by choosing a different model class or including other covariates, interaction terms or non-linear relationships [7].

One balancing diagnostic consists in evaluating the empirically standardised differences of the means of continuous covariates and, for dichotomous outcomes, the prevalence, between treated and their matched non-treated patients. These standardised differences, in case of matching with a fixed number M, are defined as

$$\begin{aligned} \kappa_{\text{continuous}} &:= \frac{\overline{x}_1 - \overline{x}_0}{\sqrt{\frac{s_1^2 + s_0^2}{2}}} \\ \kappa_{\text{dichotomous}} &:= \frac{\widehat{p}_1 - \widehat{p}_0}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1) + \widehat{p}_0(1-\widehat{p}_0)}{2}}}, \end{aligned} \tag{4.19}$$

where $\overline{x}$, s^2 and $\widehat{p}$ denote the sample means, sample variances and prevalences of the covariates

for patients in the treated and non-treated group, respectively, depending on the subscript. That is, the standardised differences measures the degree to which the mean values and prevalences differ between treated and non-treated patients, standardised by the pooled sample standard deviation. If the standardised differences are high (conventionally 0.1), this indicates that the propensity score model needs to be re-evaluated [7].

In the case of many-to-one matching with a differing number M_i of matches for each treatment group patient, we need to weight the standardised differences corresponding to the estimate in equation (4.6). Here, we will use the weighted sample mean and variances instead. Defining the weights as $a_i = 1/M_i$ and assuming that dichotomous variables are coded as 0 for non-presence and 1 for presence of the covariate, these are defined [51] by

$$\begin{aligned} \overline{x}_w &= \widehat{p}_w = \frac{\sum_{k=1}^{n_w} a_k x_k}{\sum_{k=1}^{n_w} a_k} \\ s_w^2 &= \frac{\sum_{k=1}^{n_w} a_k}{\left(\sum_{k=1}^{n_w} a_k\right)^2 - \sum_{k=1}^{n_w} a_k^2} \sum_{k=1}^{n_w} a_k (x_k - \overline{x}_w)^2. \end{aligned} \tag{4.20}$$

The condition in equation (4.1) can further be assessed by e.g. visual inspection of empirical distribution functions (for continuous covariates) or histograms (for categorical and/or continuous covarites) of the covariate in question, for the treatment and non-treatment groups, respectively.

Another important question regards which covariates should be included in the model for the propensity score. Since the propensity score is defined as $\mathbb{P}(W = 1 \mid X = x)$, there are theoretical reasons for including all covariates that are affecting the treatment variable among the pooled data. However, since the role of the propensity score is to balance confounders between treated and non-treated patients, the goal is to include the confounders and potential confounders. Potential confounders are baseline covariates suspected (e.g. from existing literature) to affect the treatment. The most important aspect in this regard is to include all confounding covariates so that the assumption of no unobserved confounders is fulfilled and to not include covariates measured after treatment assignment. However, including more than these covariates have been shown to increase the variance of the treatment estimate without reducing the bias. Furthermore, it seems that a greater number of matches are possible when only these covariates are included. However, in practice, it can be difficult to identify only the true and potential confounders, and failing to include these results in the strong ignorability assumption to not be fulfilled. In many applications, most covariates measured at patient-level are potential confounders [7].

5 Digital Twins

In this chapter we will describe another approach to using historical data when conducting clinical trials, namely the use of *digital twins*. We will focus exclusively on the use of digital twins for two-arm trials, but the method can be extended for use in single arm trials. Similar to the SCA approach the main goal is to estimate the ATE, but in this case the synthetic observations will not be directly included as observations pooled with data from the current RCT. Instead the idea behind digital twins is to use historical data to train a model which can be used to get predictions of the potential control outcome of subjects in a clinical trial and use this as an adjustment covariate in the ANCOVA model given in equation (2.20).

The main purpose of this chapter is to describe the benefits and derive useful asymptotic theoretical properties of estimating the ATE using digital twins with the ANCOVA model previously described. In order to do so, we will work with results regarding (efficient) influence functions. Therefore, we describe this general setup in section 5.2, and conclude the section by stating an important result. This result is then used in section 5.3, where we derive important theoretical properties of using digital twins. Based on some of these results, we will describe in section 5.4 how to carry out sample size calculations when using digital twins.

Throughout this chapter, we use the notation for the historical data introduced in the beginning of chapter 4. We generally assume a setup an RCT, where the randomisation scheme ensures that the W_i's of the current data points are independent, hence making the current data IID. Also, similar considerations regarding the quality of the historical data as those stated in chapter 4 suggested by Pocock [44] should be made. However, as we will later discover, any violation of these does not lead to an increase in the type I error rate as was the case for the SCA approach. However, such violations could potentially lead to smaller benefit in terms of power by using the method of digital twins.

5.1 The Digital Twins Approach

We will use the concept of digital twins described by Unlearn.AI and examine whether their use of digital twins can be theoretically justified. Specifically, we will follow Schuler et al. [35] published by Unlearn.AI, in which the theoretical benefits of using digital twins are derived. The article was peer reviewed and published during the process of writing this thesis [52]. Due to working with the unpublished paper, we went thoroughly through the proofs. Furthermore, the article contains only few details concerning the theoretical derivations, and thus we wanted to go thoroughly through the details.

In order to introduce the concept of the digital twin approach, we begin by considering some learning algorithm, which, trained on the historical dataset (X', Y'), outputs a prediction model

$m\colon \mathcal{X} \to \mathbb{R}$. Then the treatment effect β_W is estimated using the current RCT data (X_i, W_i, Y_i), e.g. from the model

$$Y_i = \beta_0 + W_i\beta_W + m(X_i)\beta_m + \varepsilon_i, \tag{5.1}$$

where $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$ are mutually independent, and we assume that W_i and X_i are mutually independent (due to randomisation) and each independent of the ε_i's.

We will seek to construct m such that $m(x_i)$ is a prediction of the expected outcome of patient i in the case they do not receive the treatment. We will refer to this artificial patient not receiving treatment as the *digital twin* of patient i. Therefore, we will refer to $m(x_i)$ as the predicted outcome of this digital twin. As specified by (5.1), the construction of digital twins are carried out for both patients receiving the treatment and for patients that are in the control group. The reason for constructing m in this way is that, as we will show (under specific conditions), the model that gives the lowest asymptotic variance of the ATE estimator is the conditional mean, $m(X_i) = \mathbb{E}[Y_i(0) \mid X_i]$. This conditional mean is referred to as the *prognostic score*, which is why we will refer to m as being a *prognostic model*. The prognostic model itself can be chosen as desired. Specifically, we will examine random, linear, LASSO and random forest prognostic models in the following chapters.

Schuler et al. [35] argue that adjusting for an estimated prognostic score is merely a formalisation of the well-established ad-hoc procedure of adjusting for prognostic covariates in clinical trials. As an example, the body mass index, the Charlson comorbidity index and the Framingham risk scores can indeed be regarded as obtained from simplified prognostic models, and the baseline measurements themselves can also be regarded as (more) rudimentary prognostic scores. In this sense, adjusting for an estimated prognostic score based on a more refined model m only provides a formalisation of prognostic covariate adjustment. Hence the model in equation (5.1) is just a special case of an ANCOVA model.

For this reason we obtain an important property of the model; we do not expect the ML estimator $\widehat{\beta}_W$ to change when introducing $m(X)$ to the regression in the setting of an RCT, but we get a more efficient estimate. This can be seen from example 2.3.3, where adjustment is with regard to $m(X)$ instead of X itself. Here, we use that X and W are independent and thus in expectation $\mathbb{X}$ and $\mathbb{W}$ are orthogonal, as well as ensuring independence between $m(X)$ and W, just by using that m is some function of X. This also ensures that adjusting for the outcome of a digital twin using the ANCOVA model does not introduce bias of the ATE estimator, even when the prognostic model is itself biased. Furthermore, under the assumption of an underlying ANCOVA model (since the model is a special case of an ANCOVA model), the ATE estimate follows an exact t-distribution, thus ensuring strict control of the type I error.

Overall, the main difference from the SCA approach is that no patients from control groups of previous clinical trials or simulated digital twins are added explicitly as new patients to the data when running the regression in (5.1). Instead, these patients are used to train the prognostic model. In this way we obtain analytically guaranteed control over the type I error rate, as opposed to the SCA approach.

However, using digital twins in single-arm studies is an SCA approach which requires the assumption that the prognostic model precisely predicts the expected outcome under the control arm, meaning that the ATE can be modelled as e.g.

$$Y_i - m(X_i) = \beta_1 + \varepsilon_i, \tag{5.2}$$

with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ being independent for each i. This is similar to the paired t-test suggested in section 4.1.1, where the external control observations are included as being paired to an internal observation from the treatment group. Using equation (5.2) the treatment is deemed effective if the null hypothesis $\beta_1 \leqslant \Delta_s$ is rejected at significance level α [53, p. 5]. However, in this case several concerns, similar to the concerns in the SCA approach, are present, one of which being that if the model does not provide an unbiased prediction of the expected value of the outcome, the ATE estimate would be biased, potentially increasing the type I error probability.

5.2 (Efficient) Influence Functions

In this section we will introduce the setup of using influence functions as a part of statistical inference. For our purposes, the aim is to use influence functions to derive asymptotic distributions of different ATE estimators. Specifically, the efficient influence function of the ATE estimand will be used to obtain the most efficient ATE estimator which turns out to be an estimator utilising digital twins. This introduction is not carried out in all details, and should only be considered as an intuitive explanation. The derivations are mostly based on [54] and [35] in which most equations are not explicitly derived. Where this is the case, we carry out the calculations, the most comprehensive of which are placed in appendix C.1.

The starting point of most statistical analyses is in modelling data with some (semi-)parametric model, which is in our case could be one of the previously specified AN(C)OVA models. In the setup of influence functions, our point of departure is nonparametric estimation. This means that we instead consider functionals of the cumulative distribution function F of the true data generating distribution (which is e.g. assumed to be normal with a linear mean structure under an AN(C)OVA model). Notice here that in general, the cumulative distribution function $F \colon \mathcal{O} \to \mathbb{R}$, where $\mathcal{O}$ is the sample space of the observations, is well defined without any assumption of a parametric model, and uniquely defines the true distribution. We can then define our estimand of interest from a functional Ψ. If e.g. the interest is in the mean of the outcome $Y \in \mathcal{O}$, we can define Ψ by

$$\Psi(F) = \mathbb{E}_F[Y] = \int y \, \mathrm{d}F = \int y f(y) \, \mathrm{d}y. \tag{5.3}$$

with the last equality holding true only when F is continuously differentiable and $f(y) = F'(y)$ for all y. If we instead want to determine the ATE from F being the true distribution of an observation $O = (X, W, Y) \in \mathcal{O}$, we have

$$\mathrm{ATE}(F) = \mathbb{E}_F[Y(1)] - \mathbb{E}_F[Y(0)] = \mathbb{E}_F\left[\mathbb{E}_F[Y(1) \mid X] - \mathbb{E}_F[Y(0) \mid X]\right], \tag{5.4}$$

where the second equality is obtained by the law of total expectation. We note that the inner expectations are with regard to the distribution of $Y \,|\, W = w, X$ for $w = 0, 1$, respectively, and the outer expectation is with regard to the distribution of X, where all distributions are in accordance with (X, W, Y) having simultaneous distribution function F.

Estimators of the above estimands $\Psi(F)$ can be obtained by substituting F by an estimator $\widehat{F}_n$ of the true distribution. If $\widehat{F}_n$ is e.g. chosen as the empirical distribution function for the observations $y_1, y_1, \ldots, y_n$, the estimator for the mean is

$$\Psi\left(\widehat{F}_n\right) = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{5.5}$$

In order to obtain an estimator for the ATE, we can replace F with F^* being any candidate distribution of (X, W, Y) such that the marginal distribution of X equals the empirical distribution of X and the conditional distribution of Y given X and W has conditional mean equal to an estimator $\widehat{\mathbb{E}}\left[Y \mid X = x, W = w\right]$. In this case we would obtain

$$\mathrm{ATE}\left(F^*\right) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathbb{E}}\left[Y \mid X = x_i, W = 1\right] - \widehat{\mathbb{E}}\left[Y \mid X = x_i, W = 0\right]. \tag{5.6}$$

Here $\widehat{\mathbb{E}}\left[Y \mid X = x, W = w\right]$ could be an estimate obtained from some model as the ANCOVA or ANOVA model.

We want to understand how sensitive $\Psi(\cdot)$ is to changes in the true distribution function F in the direction of $\widehat{F}_n$. In order to do so, we begin by regarding how $\Psi(\cdot)$ changes in the direction of a fixed deterministic distribution F^*, where the support of F^* is contained in the support of F. We then define the *parametric submodel* as the cumulative distribution function $F_t = tF^* + (1-t)F$ for $t \in [0, 1]$. Then, if the pathwise (or directional) derivative

$$\lim_{t \to 0^+}\left(\frac{\Psi(F_t) - \Psi(F)}{t}\right) = \left.\frac{\mathrm{d}\Psi(F_t)}{\mathrm{d}t}\right|_{t=0} \tag{5.7}$$

exists, we define it as the *Gâteaux derivative* of Ψ at F in the direction F^*. This derivative shares many of the same properties as the ordinary derivative, such as the chain and product rules [54]. We define a parametric submodel as *regular* if it satisfies that $F^*(O)/F(O)$ has finite variance for the stochastic variable $O \in \mathcal{O}$. This ensures that $\left.\frac{\mathrm{d}F_t}{\mathrm{d}t}\right|_{t=0}$ is well-defined [54, p. 3]. When (5.7) exists for all regular parametric submodels, we say that Ψ is *pathwise differentiable*.

According to Hines et al. [54], we define the *efficient influence function* of a pathwise differentiable estimand as a function $\varphi_F \colon \mathcal{O} \to \mathbb{R}$ that fulfills

$$\left.\frac{\mathrm{d}\Psi(F_t)}{\mathrm{d}t}\right|_{t=0} = \mathbb{E}_{F^*}\left[\varphi_F\left(O\right)\right] \tag{5.8}$$

for any F^* that makes F_t regular. The left hand side describes how sensitive the estimand is to changes of F in the direction of F^*. This implies that we can think intuitively of the efficient influence function evaluated at some point o^* stemming from the distribution F^* as describing the sensitivity of the estimand when F changes in the direction of F^*.

As already noted, if we instead of the true data distribution F use an approximated distribution (usually specified through a parametrised distribution), we would obtain an estimator of the estimand. When determining the Gâteaux derivative of this estimator and using this derivative to determine a function that satisfies equation (5.8), we would obtain the *influence function* for this estimator. Thus, the *efficient* influence function is the specific influence function of the estimand which is obtained by using the true distribution. Thus, sometimes we will exclude the subscript from φ since it will be clear from the context if it is the efficient influence function or an influence function linked to some non-true distribution.

Influence functions can be used to determine the asymptotic distribution of a *regular* and *asymptotically linear* (RAL) estimator. The definition of a RAL estimator can be found in [55]. A regular estimator $\Psi(F^*)$ of $\Psi(F)$ can be informally thought of as an estimator where small changes in F does not affect the asymptotic distribution of $\Psi(F^*)$. Regularity is a smoothness assumption on $\Psi(F^*)$ that ensures that estimators which require knowledge of the true distribution F are ruled out by requiring that the estimator is reasonable also when the true distribution is shifted by an arbitrarily small amount [31]. An asymptotically linear estimator is informally an estimator where the error between $\Psi(F^*)$ and $\Psi(F)$ for any specific n can be approximately written as a linear function with mean 0. Asymptotic linearity allows us to assess the asymptotic distribution of the estimator [31]. As it is beyond the scope of this thesis, we will not delve into further details of RAL estimators. For our purposes, it is enough to know that most reasonable estimarors are RAL, as noted in [55]. Furthermore, Tsiatis et al. [56, p. 27] note that the most efficient regular estimator is asymptotically linear [57]. For these reasons we will not show that specific estimators are indeed RAL in this thesis.

According to [54] and [56, pp. 41–42], any RAL estimator $\Psi(F^*)$ has the limiting distribution

$$\sqrt{n}\left(\Psi(F^*) - \Psi(F)\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}_F\left[\varphi_{F^*}(O)^2\right]\right), \tag{5.9}$$

where $\Psi(F^*)$ is the estimator from the distribution F^* using n observations, and where φ_{F^*} is the influence function of the estimator $\Psi(F^*)$. It can be shown that the estimator using the true cumulative distribution F obtains the lowest asymptotic variance among the class of RAL estimators [58].

5.2.1 Efficient Influence Functions of the ATE and Population Mean

In this section we wish to derive the efficient influence functions of the ATE and the population mean, which can be done since these estimands are both pathwise differentiable [54]. We do this in accordance with Hines et al. [54], meaning that we will make use of a strategy, where we utilise a cumulative distribution function $F^* = H_{o^*}$ assigning the probability 1 of a stochastic

variable taking a fixed value o^*. In order to formally define this cumulative distribution function H_{o^*}, we will need the *Dirac delta function*. For our purpose, we will first define the Dirac delta function associated to the point $o^* \in \mathbb{R}$ and evaluated at any set $A \subseteq \mathbb{R}$ as the measure

$$\mathbb{1}_{o^*}(A) = \begin{cases} 1 & \text{if } o^* \in A \\ 0 & \text{if } o^* \notin A \end{cases}. \tag{5.10}$$

Later, we will generalise the Dirac measure to the case where $o, o^* \in \mathbb{R}^k$. The *Heaviside step function* H can be expressed as the Lebesgue integral with respect to the Dirac measure, since

$$H_{o^*}(o) = \int_{-\infty}^{o} \mathrm{d}\mathbb{1}_{o^*}(u) = \mathbb{1}_{o^*}\left(]-\infty, o]\right) = \begin{cases} 1 & \text{if } o \geqslant o^* \\ 0 & \text{if } o < o^* \end{cases}. \tag{5.11}$$

We can regard the Heaviside function as the cumulative distribution function of a random variable with probability 1 of taking the value o^*. That is, we can heuristically regard $\mathbb{1}_{o^*}$ as the corresponding density function of this distribution. For some $\mathbb{1}_{o^*}(u)$-measurable function $g\colon \mathbb{R} \to \mathbb{R}$, we then have

$$\int g(u)\ \mathrm{d}\mathbb{1}_{o^*}(u) = \int g(u)\ \mathrm{d}H_{o^*}(u), \tag{5.12}$$

where the left hand side is the Lebesgue integral and the right hand side is the Riemann-Stieltjes integral. The equality is obtained directly from the definitions of the Lebesgue- and Riemann-Stieltjes integrals and using equation (5.11) to see that both sides equal $g(o^*)$. The function H is discontinuous, but had it been continuously differentiable and g continuous, the right hand side would equal $\int g(u)\mathbb{1}_{o^*}(u)\ \mathrm{d}u$. In the following, we will sometimes abuse notation and denote the integral in (5.12) in this way.

Using the parametric submodel

$$F_t(o) = tH_{o^*}(o) + (1-t)F(o), \tag{5.13}$$

the efficient influence function from equation (5.8) reduces to

$$\left.\frac{\mathrm{d}\Psi(F_t)}{\mathrm{d}t}\right|_{t=0} = \mathbb{E}_{H_{o^*}}\left[\varphi_F(O)\right] = \int \varphi_F(o)\ \mathrm{d}H_{o^*}(o) = \varphi_F(o^*). \tag{5.14}$$

Thus, using this parametric submodel, we can derive the efficient influence function evaluated in o^* of an estimand by taking the Gâteaux derivative.

For some fixed $o^* = (o_1^*, o_2^*, \ldots, o_k^*) \in \mathbb{R}^k$ and arbitrary $A = (A_1, A_2, \ldots, A_k) \subseteq \mathbb{R}^k$, the Dirac delta function is generalised by the product measure $\mathbb{1}_{o^*}(A) = \mathbb{1}_{o_1^*}(A_1)\mathbb{1}_{o_2^*}(A_2)\cdots\mathbb{1}_{o_k^*}(A_k)$. Using this generalisation, in the case of the ATE, we have that o denotes a realisation of the

stochastic vector (X, W, Y). Similar arguments as the above can be employed to obtain equivalent results. We also note that the previous calculations and following derivations are done for continuous observations o, but in the case of discrete observations or a mixed distribution, integrals can be swapped by sums, and the Dirac delta function can be replaced by an indicator function that directly specifies the probability mass function [54].

We are now ready use equation (5.14) to determine the efficient influence function of the population mean estimand in equation (5.3), which will be done in the following lemma. This lemma will be used later to prove properties of using digital twins in theorem 5.3.11.

Lemma 5.2.1.
The efficient influence function of the population mean estimand $\Psi(F)$ in equation (5.3) is given as

$$\varphi(y) = y - \Psi(F) \tag{5.15}$$

[54]. ◀

To see why this holds, we define the parametric submodel F_t correspondingly, perturbing F in the direction of the cumulative distribution of a single point o^*, which in this case is y^*. Specifically, for $t \in [0, 1]$, we use the parametric submodel F_t in equation (5.13). We then obtain

$$\begin{aligned} F_t(o) &= tH_{o^*}(o) + (1-t)F(o) = t\mathbb{1}_{o^*}\left(]-\infty, o]\right) + (1-t)F(o) \\ &= t\int_{-\infty}^{o} \mathbb{1}_{o^*}(u)\,\mathrm{d}u + (1-t)\int_{-\infty}^{o} f(u)\,\mathrm{d}u = \int_{-\infty}^{o} t\mathbb{1}_{o^*}(u) + (1-t)f(u)\,\mathrm{d}u, \end{aligned} \tag{5.16}$$

where the second to last and last equality is heuristic in the sense that the integral is with respect to the Dirac measure but we abuse notation to write it with respect to the Lebesgue measure. Using this parametric submodel with the population mean estimand and $o^* = y^*$, we get

$$\Psi(F_t) = t\int y\mathbb{1}_{y^*}(y)\,\mathrm{d}y + (1-t)\int yf(y)\,\mathrm{d}y = ty^* + (1-t)\Psi(F). \tag{5.17}$$

Taking the Gâteaux derivative, we now obtain

$$\left.\frac{\mathrm{d}\Psi(F_t)}{\mathrm{d}t}\right|_{t=0} = \lim_{t\to 0^+}\left(\frac{ty^* + (1-t)\Psi(F) - \Psi(F)}{t}\right) = y^* - \Psi(F). \tag{5.18}$$

Now using equation (5.14) we have heuristically proved the lemma.

We will now derive the efficient influence function of the ATE estimand. We will start by stating and heuristically proving lemma 5.2.2 using the same parametric submodel as in equation (5.13).

Lemma 5.2.2.
Let f denote the probability density function consistent with some cumulative distribution function F, and define the estimand $\Psi(F) = f(o)$ for some fixed o. Then the Gâteaux derivative of f at o in the direction o^*, that is, using the parametric submodel in equation (5.13), can be expressed as

$$\left.\frac{\mathrm{d}\Psi(F_t)}{\mathrm{d}t}\right|_{t=0} = \left.\frac{\mathrm{d}f_t(o)}{\mathrm{d}t}\right|_{t=0} = \mathbb{1}_{o*}(o) - f(o) \tag{5.19}$$

[54]. ◀

To see why this holds, we first note that $\Psi(F) = f(o)$ is the functional taking as input a cumulative distribution function and gives the associated density function evaluated at o. We then use the parametric submodel in equation (5.13), and follow the multidimensional equivalent steps as in (5.16) to obtain

$$\Psi(F_t) = f_t(o) = t\mathbb{1}_{o*}(o) + (1-t)f(o). \tag{5.20}$$

We then get from the definition of the Gâteaux derivative that

$$\left.\frac{\mathrm{d}\Psi(F_t)}{\mathrm{d}t}\right|_{t=0} = \lim_{t\to 0^+}\left(\frac{t\mathbb{1}_{o*}(o) + (1-t)f(o) - f(o)}{t}\right) = \mathbb{1}_{o*}(o) - f(o). \tag{5.21}$$

We can now use this lemma to determine the efficient influence function of the ATE.

Lemma 5.2.3.
The efficient influence function of the ATE estimand is $\varphi_{\mathrm{ATE}} = \varphi_1 - \varphi_0$, where

$$\begin{aligned}\varphi_1(x,w,y) &= \frac{w}{f\,(1 \mid x)}\big(y - \gamma_1(x,F)\big) + \gamma_1(x,F) - \Psi_1(F)\\ \varphi_0(x,w,y) &= \frac{1-w}{f\,(0 \mid x)}\big(y - \gamma_0(x,F)\big) + \gamma_0(x,F) - \Psi_0(F)\end{aligned} \tag{5.22}$$

and

$$\begin{aligned}\gamma_w(x,F) &= \mathbb{E}_F\left[Y \mid X = x, W = w\right]\\ \Psi_w(F) &= \mathbb{E}_F\left[\mathbb{E}_F[Y \mid X, W = w]\right]\end{aligned} \tag{5.23}$$

for $w = 0, 1$ and $(X, W, Y) \sim F$ [54, 35]. ◀

Proof. We can express the ATE as

$$\mathrm{ATE}(F) = \Psi_1(F) - \Psi_0(F). \tag{5.24}$$

We will then work with Ψ_1 and abuse notation, writing f_t as conditional and marginal distributions of X, W and Y, under a parametric submodel F_t, and similarly for the true density function f associated with F. Then, we get

$$\Psi_1(F_t) = \int\int y f_t(y \mid W = 1, X = x)\, \mathrm{d}y\; f_t(x)\, \mathrm{d}x = \int\int y \frac{f_t(x,1,y) f_t(x)}{f_t(1,x)}\, \mathrm{d}y\, \mathrm{d}x. \tag{5.25}$$

Now, we wish to evaluate $\frac{\mathrm{d}\Psi_1(F_t)}{\mathrm{d}t}\big|_{t=0}$. Here, we specify the parametric submodel as in equation (5.13). Having done so, and using equation (5.14), we thus obtain an expression of the influence function as the Gâteaux derivative. Assuming regularity conditions (so that the order of integration and Gâteaux derivative can be interchanged) and using lemma 5.2.2, the product- and chain rules of the Gâteaux derivative and properties of the Dirac delta function, we obtain

$$\left.\frac{\mathrm{d}\Psi_1(F_t)}{\mathrm{d}t}\right|_{t=0} = \frac{\mathbb{1}_{w*}(1)}{f\,(1 \mid x^*)}\left(y^* - \gamma_1(x^*, F)\right) + \gamma_1(x^*, F) - \Psi_1(F) = \varphi_1(x^*, w^*, y^*), \tag{5.26}$$

which we derive in further details in appendix C.1. These derivations can also be carried out exchanging $W = 1$ with $W = 0$, obtaining that

$$\left.\frac{\mathrm{d}\Psi_0(F_t)}{\mathrm{d}t}\right|_{t=0} = \frac{\mathbb{1}_{w*}(0)}{f\,(0 \mid x^*)}\left(y^* - \gamma_0(x^*, F)\right) + \gamma_0(x^*, F) - \Psi_0(F) = \varphi_0(x^*, w^*, y^*). \tag{5.27}$$

Using linearity of the Gâteaux derivative, we then get that the efficient influence function of the ATE evaluated in (x^*, w^*, y^*) can be expressed as

$$\varphi_{\mathrm{ATE}}(x^*, w^*, y^*) = \left.\frac{\mathrm{d}\Psi_1(F_t)}{\mathrm{d}t}\right|_{t=0} - \left.\frac{\mathrm{d}\Psi_0(F_t)}{\mathrm{d}t}\right|_{t=0} = \varphi_1(x^*, w^*, y^*) - \varphi_0(x^*, w^*, y^*). \tag{5.28}$$

Hence, taking the efficient influence function with a general input $(x, w, y) \in \mathcal{X} \times \{0, 1\} \times \mathbb{R}$, we get the result. ■

Note that the lemma holds in general, even when the assumptions of an RCT are not met, as we have assumed for this whole chapter. Using the lemma, we can directly get an expression of the efficient influence function of the ATE estimand when in the case of an RCT that will be useful in proving statements later in the chapter. Furthermore, in the rest of the chapter, we will use that $W \perp\!\!\!\perp X$ in an RCT, and hence $f(w \mid x) = f(w)$. Therefore as in section 2.2.1 we can define

$$f(1) = \mathbb{E}[W] = \pi_1, \quad f(0) = \mathbb{E}[1 - W] = \pi_0. \tag{5.29}$$

Corollary 5.2.4.
The efficient influence function of the ATE estimand can be written as

$$\begin{aligned}\varphi_{\mathrm{ATE}}(X, W, Y) = {} & \frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1)]\right) - \frac{1 - W}{\pi_0}\left(Y - \mathbb{E}[Y(0)]\right) \\ & - \frac{W - \pi_1}{\pi_0 \pi_1}\Big(\pi_1\left(\mathbb{E}[Y(0) \mid X] - \mathbb{E}[Y(0)]\right) + \pi_0\left(\mathbb{E}[Y(1) \mid X] - \mathbb{E}[Y(1)]\right)\Big),\end{aligned} \tag{5.30}$$

where the relevant expectations are taken with regard to the true distribution F [54, 35]. ◀

Proof. By similar arguments as in (2.19), we have that $\gamma_w = \mathbb{E}\left[Y \mid X, W = w\right] = \mathbb{E}[Y(w) \mid X]$ in the case of an RCT. By lemma 5.2.3, we can then express the efficient influence function of the ATE estimand as

$$\begin{aligned}
\varphi_{\mathrm{ATE}}(X, W, Y) = &\frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1) \mid X]\right) + \mathbb{E}[Y(1) \mid X] - \mathbb{E}\left[\mathbb{E}[Y(1) \mid X]\right] \\
&- \left(\frac{1-W}{\pi_0}\left(Y - \mathbb{E}[Y(0) \mid X]\right) + \mathbb{E}[Y(0) \mid X] - \mathbb{E}\left[\mathbb{E}[Y(0) \mid X]\right]\right) \\
= &\frac{W}{\pi_1} Y - \mathbb{E}[Y(1)] - \frac{W}{\pi_1}\mathbb{E}[Y(1) \mid X] + \mathbb{E}[Y(1) \mid X] \\
&- \frac{1-W}{\pi_0} Y + \mathbb{E}[Y(0)] + \frac{1-W}{\pi_0}\mathbb{E}[Y(0) \mid X] - \mathbb{E}[Y(0) \mid X] \qquad (5.31) \\
= &\frac{W}{\pi_1} Y - \frac{W}{\pi_1}\mathbb{E}[Y(1)] + \frac{W-\pi_1}{\pi_1}\mathbb{E}[Y(1)] - \frac{W-\pi_1}{\pi_1}\mathbb{E}[Y(1) \mid X] \\
&- \frac{1-W}{\pi_0} Y + \frac{1-W}{\pi_0}\mathbb{E}[Y(0)] + \frac{W-\pi_1}{\pi_0}\mathbb{E}[Y(0)] - \frac{W-\pi_1}{\pi_0}\mathbb{E}[Y(0) \mid X] \\
= &\frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1)]\right) - \frac{1-W}{\pi_0}\left(Y - \mathbb{E}[Y(0)]\right) \\
&- \frac{W-\pi_1}{\pi_0\pi_1}\left(\pi_1\left(\mathbb{E}[Y(0) \mid X] - \mathbb{E}[Y(0)]\right) + \pi_0\left(\mathbb{E}[Y(1) \mid X] - \mathbb{E}[Y(1)]\right)\right),
\end{aligned}$$

where we use in the second equality that $\mathbb{E}\left[\mathbb{E}[Y(W) \mid X]\right] = \mathbb{E}[Y(W)]$ due to the law of total expectation, and the third equality follows from

$$\mathbb{E}[Y(0)] = \frac{1-(1-\pi_0)}{\pi_0}\mathbb{E}[Y(0)] = \frac{1-\pi_1}{\pi_0}\mathbb{E}[Y(0)] = \frac{1-W}{\pi_0}\mathbb{E}[Y(0)] + \frac{W-\pi_1}{\pi_0}\mathbb{E}[Y(0)],$$

and

$$\frac{1-W}{\pi_0}\mathbb{E}[Y(0) \mid X] - \mathbb{E}[Y(0) \mid X] = \frac{(1-W)-\pi_0}{\pi_0}\mathbb{E}[Y(0) \mid X] = -\frac{W-\pi_1}{\pi_0}\mathbb{E}[Y(0) \mid X],$$

thus completing the proof. ■

Directly from lemma 5.2.3, we can show that the influence function of any ATE estimator has mean 0, which will be useful later when proving theorem 5.3.11.

Corollary 5.2.5.
Under any assumed distribution F^*, the influence function of the corresponding ATE estimator has mean 0. ◀

Proof. The influence function of the ATE estimator $\mathrm{ATE}(F^*)$ is given by corollary 5.2.4, assuming $(X, W, Y) \sim F^*$. Thus, we get that

$$\begin{aligned}
\mathbb{E}_{F^*}[\varphi_{\mathrm{ATE}}] &= \mathbb{E}_{F^*}\left[\frac{W}{\pi_1}\left(Y - \mathbb{E}_{F^*}[\gamma_1(X, F^*)]\right) - \frac{1-W}{\pi_0}\left(Y - \mathbb{E}_{F^*}[\gamma_0(X, F^*)]\right)\right] \\
&= \frac{1}{\pi_1}\left(\mathbb{E}_{F^*}[WY] - \mathbb{E}_{F^*}[W]\,\mathbb{E}_{F^*}[\gamma_1(X, F^*)]\right) \\
&\quad - \frac{1}{\pi_0}\left(\mathbb{E}_{F^*}[(1-W)Y] - \mathbb{E}_{F^*}[1-W]\,\mathbb{E}_{F^*}[\gamma_0(X, F^*)]\right) \\
&= \frac{\mathbb{E}_{F^*}[W]}{\pi_1}\left(\mathbb{E}_{F^*}[Y(1)] - \mathbb{E}_{F^*}[\gamma_1(X, F^*)]\right) \\
&\quad - \frac{\mathbb{E}_{F^*}[1-W]}{\pi_0}\left(\mathbb{E}_{F^*}[Y(0)] - \mathbb{E}_{F^*}[\gamma_0(X, F^*)]\right) = 0,
\end{aligned} \tag{5.32}$$

where we use in the first equality that $\mathbb{E}_{F^*}[\gamma_w(X, F^*)] = \Psi_w(F^*) = \mathbb{E}_{F^*}[\Psi_w(F^*)]$. In the third equality, we use the representation $Y = WY(1) + (1-W)Y(0)$ and that W is independent of potential outcomes. In the last equality, we use the law of total expectation to see that $\mathbb{E}_{F^*}[\gamma_w(X, F^*)] = \mathbb{E}_{F^*}\left[\mathbb{E}_{F^*}\left[Y(w) \mid X\right]\right] = \mathbb{E}_{F^*}[Y(w)]$. ■

5.3 Theoretical Properties of the Digital Twins Approach

In the following, we will consider three different AN(C)OVA models to estimate the ATE. Firstly, we will consider the *difference-in-means* model given by

$$Y = \beta_0 + W\beta_W + \varepsilon_W \tag{5.33}$$

where W is independent of both $\varepsilon_1 \sim \mathcal{N}_{n_1}\left(0, \sigma_1^2 I_{n_1}\right)$ and $\varepsilon_0 \sim \mathcal{N}_{n_0}\left(0, \sigma_0^2 I_{n_0}\right)$, which are also independent from each other. This model differs from the ANOVA model by allowing different variances in the treatment and control group, where $\sigma_0^2 = \mathbb{V}\mathrm{ar}\left(Y(0)\right)$ and $\sigma_1^2 = \mathbb{V}\mathrm{ar}\left(Y(1)\right)$ from the model specification. Under this model, we have the difference-in-means ATE estimator given by

$$\widehat{\mathrm{ATE}}_{\Delta} = \widehat{\beta}_W = \overline{Y}_1 - \overline{Y}_0, \tag{5.34}$$

which coincides with the usual ANOVA estimator. Secondly, we will refer to the ANCOVA model introduced in (2.20) with design matrix

$$\mathbb{D} = [1\ \mathbb{W}\ \mathbb{X}] \tag{5.35}$$

as the *ANCOVA I model*. As described earlier, an estimate of the ATE can be obtained as $\widehat{\beta}_W$ from this model. Thirdly, we consider an ANCOVA model where the covariates $\mathbb{X}$ are demeaned, denoted as $\widetilde{\mathbb{X}} = \mathbb{X} - \mathbb{E}[\mathbb{X}]$. In this model, we will include interaction effects between the covariates and the treatment allocation, obtaining the design matrix

$$\mathbb{D} = \begin{bmatrix} 1 & \mathbb{W} & \widetilde{\mathbb{X}} & \operatorname{diag}_n(\mathbb{W})\widetilde{\mathbb{X}} \end{bmatrix}. \tag{5.36}$$

We will refer to this model as the *ANCOVA II model*. The model assumptions is then that the data is on the form

$$Y_i = \beta_0 + W_i\beta_W + \widetilde{X}_i\beta_{\widetilde{X}} + W_i\widetilde{X}_i\beta_{W\times\widetilde{X}} + \varepsilon_i, \tag{5.37}$$

where $\varepsilon_i \sim \mathcal{N}_n\left(0, \sigma^2\right)$ and the ε_i's are independent. The ATE can be estimated from this model as $\widehat{\beta}_W$, since

$$\begin{aligned} \text{ATE} &= \mathbb{E}\left[Y(1) - Y(0)\right] \\ &= \mathbb{E}\left[\beta_W + \widetilde{X}\beta_{W\times\widetilde{X}}\right] = \beta_W, \end{aligned} \tag{5.38}$$

which is obtained by $\mathbb{E}[\widetilde{X}] = 0$, since the covariates are demeaned. Intuitively, the ATE is β_W when we demean X, since β_W is the treatment effect of the patient with $\widetilde{X} = 0$, which is the patient with X being equal to the population mean, that is, the "average patient". The purpose of including interactions is solely to decrease the residual variance σ^2 compared to an ANCOVA model without interaction effects. We note that this decrease in σ^2 will only be present if there is in fact heterogeneity of the treatment effect across the levels of the covariates, since in the opposite case, the interaction effects $\beta_{W\times\widetilde{X}} = 0$. In this situation, when including interaction effects we would decrease the model degrees of freedom by p, hence yielding an increase in $\widehat{\sigma}^2$. However, in the following, we will consider the asymptotic variance. As we will argue, using the ANCOVA I ATE estimator with digital twins provides the most asymptotically efficient estimate under constant treatment effect among all RAL estimators, while the ANCOVA II estimator should be used when a heterogeneous effect is present.

In the case of estimating the ATE using the ANCOVA I model, we implicitly assume that the data generating process has the cumulative distribution function F_I which describes the distribution given by the model specification in equation (2.21). Furthermore, since we are in the case of an RCT, we have that $Y(w) \mid X$ and $Y \mid X, W = w$ have the same distribution. Therefore, according to lemma 5.2.3, we obtain

$$\begin{aligned} \gamma_w(x, F_I) &= \beta_0 + w\beta_W + x\beta_X \\ \Psi_w(F_I) &= \beta_0 + w\beta_W + \mathbb{E}[X]\beta_X. \end{aligned} \tag{5.39}$$

Similarly, denoting by F_{II} the distribution under the ANCOVA II model, we get that

$$\begin{aligned} \gamma_w(x, F_{II}) &= \beta_0 + w\beta_W + \widetilde{x}\beta_{\widetilde{X}} + w\widetilde{x}\beta_{W\times\widetilde{X}} \\ \Psi_w(F_{II}) &= \beta_0 + w\beta_W + \mathbb{E}\left[\widetilde{X}\right]\beta_{\widetilde{X}} + w\,\mathbb{E}\left[\widetilde{X}\right]\beta_{W\times\widetilde{X}}. \end{aligned} \tag{5.40}$$

We note that for this latter model specification, $(\widetilde{X}, \widetilde{X}W)$ plays the role of the X mentioned in the lemma.

We can use these equations to obtain the influence function of the ATE under the ANCOVA I and II models. We emphasize the difference between the *efficient* influence function using the true distribution F and an influence function using an assumed distribution in place of F, e.g. specified by the above AN(C)OVA model specifications.

The derivations in the following subsections are based on Schuler et al. [35], in which most equations are not explicitly derived. Where this is the case, we carry out the calculations, the most comprehensive of which are placed in appendices C.2 – C.7.

5.3.1 Asymptotic Distributions of AN(C)OVA Estimators

In this section, we will derive asymptotic distributions of the difference-in-means, ANCOVA I and ANCOVA II ATE estimators, as well as relating these to each other. First, we prove a lemma, which will enable us to show consistency of the ANCOVA estimators.

Lemma 5.3.1.
Assume that for an estimator $\widehat{\theta}_n$ of θ, $\mathbb{V}\mathrm{ar}(\widehat{\theta}_n) \longrightarrow 0$ and $\mathbb{E}[\widehat{\theta}_n] \longrightarrow \theta$ as $n \to \infty$. Then the estimator is consistent, that is, $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta$. ◀

Proof. Fix $\varepsilon > 0$. From the convergence of the expected value, we have for a large enough n that

$$\left|\mathbb{E}[\widehat{\theta}_n] - \theta\right| < \varepsilon/2 \tag{5.41}$$

Thus, using Chebyshev's inequality, we have for large enough n that

$$\begin{aligned}\mathbb{P}\left(\left|\widehat{\theta}_n - \theta\right| > \varepsilon\right) &\leqslant \mathbb{P}\left(\left|\widehat{\theta}_n - \mathbb{E}[\widehat{\theta}_n]\right| + \left|\mathbb{E}[\widehat{\theta}_n] - \theta\right| > \varepsilon\right) \leqslant \mathbb{P}\left(\left|\widehat{\theta}_n - \mathbb{E}[\widehat{\theta}_n]\right| > \varepsilon/2\right) \\ &\leqslant \frac{\mathbb{V}\mathrm{ar}(\widehat{\theta}_n)}{\varepsilon^2/4},\end{aligned} \tag{5.42}$$

which converges to 0 as $n \to \infty$, showing that $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta$. ■

Lemma 5.3.2.
The difference-in-means ATE estimator is consistent with $\sqrt{n}\left(\widehat{\mathrm{ATE}}_\Delta - \mathrm{ATE}\right) \xrightarrow{\mathbb{P}} 0$ and has asymptotic variance

$$n\,\mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_\Delta\right) \longrightarrow \frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1}, \tag{5.43}$$

where $\sigma_w^2 = \mathbb{V}\mathrm{ar}\left(Y(w)\right)$. ◀

Proof. We see from corollary 5.2.4 that

$$\varphi_\Delta(W,Y) = \frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1)]\right) - \frac{1-W}{\pi_0}\left(Y - \mathbb{E}[Y(0)]\right) \tag{5.44}$$

is the influence function for $\widehat{\mathrm{ATE}}_\Delta$, since the last term in equation (5.30) is 0 because we do not adjust by any X under the difference-in-means model. Thus using the result in (5.9), $\widehat{\mathrm{ATE}}_\Delta$ has the limiting distribution

$$\sqrt{n}\left(\widehat{\mathrm{ATE}}_\Delta - \mathrm{ATE}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[\varphi_\Delta^2\right]\right). \tag{5.45}$$

The limiting variance can thus be expressed as

$$\begin{aligned}
\mathbb{E}\left[\varphi_\Delta^2\right] &= \mathbb{E}\left[\left(\frac{W}{\pi_1}\left(Y(1) - \mathbb{E}\left[Y(1)\right]\right) - \frac{1-W}{\pi_0}\left(Y(0) - \mathbb{E}\left[Y(0)\right]\right)\right)^2\right] \\
&= \mathbb{E}\left[\frac{W}{\pi_1^2}\left(Y(1) - \mathbb{E}\left[Y(1)\right]\right)^2 + \frac{1-W}{\pi_0^2}\left(Y(0) - \mathbb{E}\left[Y(0)\right]\right)^2\right] \\
&= \frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1}.
\end{aligned} \tag{5.46}$$

Now since the ATE estimator converges in distribution, the moments converge towards the moment parameters in the distribution [59, p. 18]. Thus we obtain the limiting variance and by lemma 5.3.1 we now have consistency of the estimator. ■

We notice that the difference-in-means ATE estimator and the ANOVA model ATE estimator have the same (asymptotic) variance when $\sigma_0 = \sigma_1$.

In the following, we use the convention that for two general row vectors U and V, being $1 \times p$ and $1 \times q$ dimensional, respectively, we have

$$\mathbb{C}\mathrm{ov}\left(U,V\right) = \mathbb{E}\left[\left(U - \mathbb{E}[U]\right)^\top \left(V - \mathbb{E}[V]\right)\right] \tag{5.47}$$

being $p \times q$ dimensional. When U and V are instead column vectors, being $p \times 1$ and $q \times 1$ dimensional, respectively, we define

$$\mathbb{C}\mathrm{ov}\left(U,V\right) = \mathbb{E}\left[\left(U - \mathbb{E}[U]\right)\left(V - \mathbb{E}[V]\right)^\top\right], \tag{5.48}$$

which is also $p \times q$ dimensional. In the following theorem, we will frequently encounter the covariance between the stochastic row vector X and the stochastic variable Y, where we will

write $\mathbb{C}\text{ov}(Y, X)$ to obtain a row vector of covariances between the response and each covariate, and $\mathbb{C}\text{ov}(X^\top, Y)$ to obtain an equivalent column vector. Note furthermore that throughout the rest of this chapter, all variances and covariances involving the covariates can be written using the non-demeaned covariates or demeaned covariates, since e.g.

$$\mathbb{C}\text{ov}(Y, X) = \mathbb{E}[YX] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}\left[Y\left(X - \mathbb{E}[X]\right)\right] = \mathbb{E}\left[Y\widetilde{X}\right] = \mathbb{C}\text{ov}\left(Y, \widetilde{X}\right).$$

With this in mind, we can now determine the asymptotic distribution of the ANCOVA I ATE estimator.

Theorem 5.3.3.
The ANCOVA I ATE estimator $\widehat{\text{ATE}}_I$ is consistent with $\sqrt{n}\left(\widehat{\text{ATE}}_I - \text{ATE}\right) \xrightarrow{\mathbb{P}} 0$ and has asymptotic variance given by

$$n\,\mathbb{V}\text{ar}\left(\widehat{\text{ATE}}_I\right) \longrightarrow \frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} + \frac{1}{\pi_0\pi_1}\xi V \xi^\top - 2\frac{1}{\pi_0\pi_1}\xi_* V \xi^\top \tag{5.49}$$

for $n \to \infty$, where

$$\begin{aligned} \sigma_w^2 &= \mathbb{V}\text{ar}\left(Y(w)\right) \\ V &= \mathbb{V}\text{ar}(X)^{-1} \\ \xi &= \mathbb{C}\text{ov}\left(Y, X\right) \\ \xi_* &= \pi_0\,\mathbb{C}\text{ov}\left(Y(1), X\right) + \pi_1\,\mathbb{C}\text{ov}\left(Y(0), X\right). \end{aligned} \tag{5.50}$$

◀

Proof. We begin by noting that under the assumed distribution F_I in equation (2.20) we have

$$Y = D\beta + \varepsilon. \tag{5.51}$$

Here, we denote by $D = (1, W, X)$ a single stochastic row of the design matrix from equation (5.35), so that X corresponds to a stochastic row vector of covariates. In the remainder of this proof, all expectations are taken with respect to this distribution F_I. Multiplying (5.51) with $D^\top$ and taking the expected value, we get that the true parameter vector β can be expressed as

$$\beta = \mathbb{E}[D^\top D]^{-1}\,\mathbb{E}[D^\top Y], \tag{5.52}$$

where we use the assumed independence between D and ε. We can now determine each of the factors in the product, starting with $\mathbb{E}[D^\top Y]$. We note that

$$WY = W\left(WY(1) + (1 - W)Y(0)\right) = WY(1), \tag{5.53}$$

so

$$\mathbb{E}\left[D^\top Y\right] = \mathbb{E}\left[\begin{pmatrix} Y \\ WY \\ X^\top Y \end{pmatrix}\right] = \begin{pmatrix} \mathbb{E}[Y] \\ \mathbb{E}[WY(1)] \\ \mathbb{E}[X^\top Y] \end{pmatrix} = \begin{pmatrix} \mathbb{E}[Y] \\ \mathbb{E}[W]\,\mathbb{E}[Y(1)] \\ \mathbb{C}\mathrm{ov}(X^\top, Y) + \mathbb{E}[X^\top]\,\mathbb{E}[Y] \end{pmatrix}, \tag{5.54}$$

where we use that in an RCT, we have the independence in (2.4). In appendix C.2.1 we show that the first factor $\mathbb{E}[D^\top D]^{-1}$ is equal to

$$\mathbb{E}\left[D^\top D\right]^{-1} = \begin{bmatrix} \frac{1}{1-\mathbb{E}[W]} + \mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{E}[X]^\top & -\frac{1}{1-\mathbb{E}[W]} & -\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}[X]^{-1} \\ -\frac{1}{1-\mathbb{E}[W]} & \frac{1}{\mathbb{E}[1-W]\,\mathbb{E}[W]} & 0 \\ -\left(\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\right)^\top & 0 & \mathbb{V}\mathrm{ar}(X)^{-1} \end{bmatrix}. \tag{5.55}$$

By multiplying these expressions, which we do in appendix C.2.2, we get that

$$\beta = \begin{pmatrix} \mathbb{E}[Y(0)] - \mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top, Y) \\ \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ \mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top, Y) \end{pmatrix}. \tag{5.56}$$

We see that all but the first entry do not depend on whether we demean X or not. In general for linear models, we can subtract the empirical mean of the covariates without affecting the MLE of the non-intercept parameters and their variances [60]. In the following we will derive the influence function of the estimator using demeaned covariates to simplify calculations, thus also deriving the asymptotic variance of the estimator not using demeaned covariates. If we demean the covariate values X, using instead $\widetilde{X}$, the first entry reduces to just $\mathbb{E}[Y(0)]$, and we obtain

$$\begin{aligned} \gamma_w(\widetilde{X}, F_I) &= \mathbb{E}[Y \mid \widetilde{X}, W = w] \\ &= (1, w, \widetilde{X})\beta \\ &= \mathbb{E}[Y(0)] + w\left(\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]\right) + \widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top, Y) \\ &= \mathbb{E}[Y(0)] + w \cdot \mathrm{ATE} + \widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top, Y). \end{aligned} \tag{5.57}$$

Taking expectation with respect to $\widetilde{X}$ in this equation, we obtain $\Psi_w(F_I)$. Now using lemma 5.2.3 together with the obtained β vector, we obtain

$$\begin{aligned} \varphi_1(\widetilde{X}, W, Y) = \frac{W}{\pi_1}\left(Y - \left(\mathbb{E}[Y(0)] + \mathrm{ATE} + \widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top, Y)\right)\right) \\ + \mathbb{E}[Y(0)] + \mathrm{ATE} + \widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top, Y) \\ - \left(\mathbb{E}[Y(0)] + \mathrm{ATE} + \mathbb{E}[\widetilde{X}]\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top, Y)\right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{W}{\pi_1}\left(Y - \left(\mathbb{E}[Y(0)] + \text{ATE} + \widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}(X^\top, Y)\right)\right) \\
&\quad + \widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}(X^\top, Y) \\
&= \frac{W}{\pi_1}\left(Y - \left(\mathbb{E}[Y(1)] + \widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}(X^\top, Y)\right)\right) \\
&\quad + \widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}(X^\top, Y) \\
&= \frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1)]\right) - \frac{W-\pi_1}{\pi_1}\widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}(X^\top, Y),
\end{aligned}
\tag{5.58}
$$

using in the second equality that $\mathbb{E}[\widetilde{X}] = 0$. In a similar way, we get

$$
\begin{aligned}
\varphi_0(\widetilde{X}, W, Y) &= \frac{1-W}{\pi_0}\left(Y - \left(\mathbb{E}[Y(0)] + \widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}(X^\top, Y)\right)\right) \\
&\quad + \mathbb{E}[Y(0)] + \widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}(X^\top, Y) \\
&\quad - \left(\mathbb{E}[Y(0)] + \mathbb{E}[\widetilde{X}]\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}(X^\top, Y)\right) \\
&= \frac{1-W}{\pi_0}\left(Y - \left(\mathbb{E}[Y(0)] + \widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}(X^\top, Y)\right)\right) \\
&\quad + \widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}(X^\top, Y) \\
&= \frac{1-W}{\pi_0}\left(Y - \mathbb{E}[Y(0)]\right) - \frac{(1-W)-\pi_0}{\pi_0}\widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}(X^\top, Y).
\end{aligned}
\tag{5.59}
$$

Taking the difference, we get

$$
\begin{aligned}
\varphi_{\text{ATE},I}(\widetilde{X}, W, Y) &= \varphi_1(\widetilde{X}, W, Y) - \varphi_0(\widetilde{X}, W, Y) \\
&= \frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1)]\right) - \frac{1-W}{\pi_0}\left(Y - \mathbb{E}[Y(0)]\right) \\
&\quad - \left(\frac{W-\pi_1}{\pi_1} - \frac{(1-W)-\pi_0}{\pi_0}\right)\widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}(X^\top, Y) \\
&= \frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1)]\right) - \frac{1-W}{\pi_0}\left(Y - \mathbb{E}[Y(0)]\right) \\
&\quad - \frac{W-\pi_1}{\pi_0\pi_1}\widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}(X^\top, Y) \\
&\coloneqq \varphi_\Delta(W, Y) - \omega(X, W, Y),
\end{aligned}
\tag{5.60}
$$

where we use the influence function of the difference-in-means estimator in equation (5.44) and we define

$$
\omega(\widetilde{X}, W, Y) = \frac{W-\pi_1}{\pi_0\pi_1}\widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}(X^\top, Y), \tag{5.61}
$$

and the third equality holds since

$$\begin{aligned}\frac{W-\pi_1}{\pi_1} - \frac{(1-W)-\pi_0}{\pi_0} &= \frac{W\pi_0 - \pi_1\pi_0}{\pi_1\pi_0} - \frac{\pi_1 - W\pi_1 - \pi_0\pi_1}{\pi_0\pi_1} \\ &= \frac{W\pi_0 - \pi_1 + W\pi_1}{\pi_1\pi_0} = \frac{W(\pi_0+\pi_1)-\pi_1}{\pi_1\pi_0} = \frac{W-\pi_1}{\pi_1\pi_0}.\end{aligned} \tag{5.62}$$

Using the result in (5.9), $\widehat{\mathrm{ATE}}_I$ has the limiting distribution

$$\sqrt{n}\left(\widehat{\mathrm{ATE}}_I - \mathrm{ATE}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[\varphi^2_{\mathrm{ATE},I}\right]\right). \tag{5.63}$$

where we drop the argument $(\widetilde{X}, W, Y)$ for the rest of this proof for ease of notation. We now wish to derive an explicit expression of this asymptotic variance. We will use that

$$\mathbb{E}\left[\varphi^2_{\mathrm{ATE},I}\right] = \mathbb{E}\left[(\varphi_\Delta - \omega)^2\right] = \mathbb{E}\left[\varphi^2_\Delta\right] + \mathbb{E}\left[\omega^2\right] - 2\,\mathbb{E}\left[\varphi_\Delta\omega\right]. \tag{5.64}$$

Using equation (5.46) we have the first term. For the second term in (5.64), we get

$$\begin{aligned}\mathbb{E}\left[\omega^2\right] &= \mathbb{E}\left[\left(\frac{W-\pi_1}{\pi_0\pi_1}\widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top, Y)\right)^2\right] \\ &= \mathbb{E}\left[\frac{(W-\pi_1)^2}{\pi_0^2\pi_1^2}\left(\widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top, Y)\right)^\top \widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top, Y)\right] \\ &= \mathbb{E}\left[\frac{(W-\pi_1)^2}{\pi_0^2\pi_1^2}\,\mathbb{C}\mathrm{ov}(Y, X)\,\mathbb{V}\mathrm{ar}(X)^{-1}\widetilde{X}^\top\widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top, Y)\right] \\ &= \frac{\mathbb{E}\left[(W-\pi_1)^2\right]}{\pi_0^2\pi_1^2}\,\mathbb{C}\mathrm{ov}(Y, X)\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{E}\left[\widetilde{X}^\top\widetilde{X}\right]\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top, Y) \\ &= \frac{1}{\pi_0\pi_1}\,\mathbb{C}\mathrm{ov}(Y, X)\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top, Y),\end{aligned} \tag{5.65}$$

where we use in the last equality that since $\widetilde{X}$ is demeaned, $\mathbb{E}\left[\widetilde{X}^\top\widetilde{X}\right] = \mathbb{V}\mathrm{ar}(X)$, and that

$$\mathbb{E}\left[(W-\pi_1)^2\right] = \mathbb{E}[W^2 + \pi_1^2 - 2\pi_1 W] = \pi_1 + \pi_1^2 - 2\pi_1^2 = \pi_1 - \pi_1^2 = \pi_1(1-\pi_1) = \pi_1\pi_0,$$

using $W^2 = W$. For the last term in equation (5.64), we have

$$\mathbb{E}\left[\varphi_\Delta\omega\right] = \mathbb{E}\left[\frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1)]\right)\omega\right] - \mathbb{E}\left[\frac{1-W}{\pi_0}\left(Y - \mathbb{E}[Y(0)]\right)\omega\right]. \tag{5.66}$$

Beginning with first term in equation (5.66), we get

$$\begin{aligned}\mathbb{E}\left[\frac{W}{\pi_1}\left(Y-\mathbb{E}[Y(1)]\right)\omega\right] &= \mathbb{E}\left[\frac{W}{\pi_1}\left(Y-\mathbb{E}[Y(1)]\right)\frac{W-\pi_1}{\pi_0\pi_1}\widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top,Y)\right]\\ &= \frac{1}{\pi_1}\mathbb{E}\left[W\left(Y-\mathbb{E}[Y(1)]\right)\frac{W-\pi_1}{\pi_0\pi_1}\widetilde{X}\right]\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top,Y)\\ &= \frac{1}{\pi_1}\,\mathbb{C}\mathrm{ov}\left(Y(1),X\right)\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top,Y), \end{aligned} \tag{5.67}$$

with the last equality following from

$$\begin{aligned}\mathbb{E}\left[W\left(Y-\mathbb{E}[Y(1)]\right)\frac{W-\pi_1}{\pi_0\pi_1}\widetilde{X}\right] &= \mathbb{E}\left[W\left(WY(1)-(1-W)Y(0)-\mathbb{E}[Y(1)]\right)\frac{W-\pi_1}{\pi_0\pi_1}\widetilde{X}\right]\\ &= \mathbb{E}\left[\left(Y(1)-\mathbb{E}[Y(1)]\right)\frac{W(W-\pi_1)}{\pi_0\pi_1}\widetilde{X}\right]\\ &= \mathbb{E}\left[\frac{W(W-1+\pi_0)}{\pi_0\pi_1}\right]\mathbb{E}\left[\left(Y(1)-\mathbb{E}[Y(1)]\right)\widetilde{X}\right]\\ &= \mathbb{E}\left[\frac{W\pi_0}{\pi_0\pi_1}\right]\mathbb{E}\left[\left(Y(1)-\mathbb{E}[Y(1)]\right)\widetilde{X}\right]\\ &= \mathbb{C}\mathrm{ov}\left(Y(1),X\right). \end{aligned} \tag{5.68}$$

The second term in equation (5.66) can be derived using equivalent calculations, yielding

$$\mathbb{E}\left[\frac{1-W}{\pi_0}\left(Y-\mathbb{E}[Y(0)]\right)\omega\right] = -\frac{1}{\pi_0}\,\mathbb{C}\mathrm{ov}\left(Y(0),X\right)\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top,Y). \tag{5.69}$$

We thus get

$$\begin{aligned}\mathbb{E}\left[\varphi_\Delta\omega\right] &= \frac{1}{\pi_1}\,\mathbb{C}\mathrm{ov}\left(Y(1),X\right)\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top,Y)\\ &\quad+\frac{1}{\pi_0}\,\mathbb{C}\mathrm{ov}\left(Y(0),X\right)\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top,Y)\\ &= \frac{1}{\pi_0\pi_1}\Big(\pi_0\,\mathbb{C}\mathrm{ov}\left(Y(1),X\right)+\pi_1\,\mathbb{C}\mathrm{ov}\left(Y(0),X\right)\Big)\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top,Y). \end{aligned} \tag{5.70}$$

Now, assembling these derivations according to (5.64), we obtain the limiting variance as

$$\begin{aligned}\mathbb{E}\left[\varphi^2_{\mathrm{ATE},I}\right] &= \frac{\sigma_0^2}{\pi_0}+\frac{\sigma_1^2}{\pi_1}+\frac{1}{\pi_0\pi_1}\,\mathbb{C}\mathrm{ov}(Y,X)\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top,Y)\\ &\quad-2\frac{1}{\pi_0\pi_1}\Big(\pi_0\,\mathbb{C}\mathrm{ov}\left(Y(1),X\right)+\pi_1\,\mathbb{C}\mathrm{ov}\left(Y(0),X\right)\Big)\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top,Y)\\ &= \frac{\sigma_0^2}{\pi_0}+\frac{\sigma_1^2}{\pi_1}+\frac{1}{\pi_0\pi_1}\xi V\xi^\top-2\frac{1}{\pi_0\pi_1}\xi_* V\xi^\top. \end{aligned} \tag{5.71}$$

Now since the ATE estimator converges in distribution, the moments converge towards the moment parameters in the distribution [59, p. 18]. Thus we obtain the limiting variance and by lemma 5.3.1 we now have consistency of the estimator. ■

Now we observe that

$$\begin{aligned}\xi &= \mathbb{C}\text{ov}\left(Y, X\right) = \mathbb{C}\text{ov}\left(WY(1), \widetilde{X}\right) + \mathbb{C}\text{ov}\left((1-W)Y(0), \widetilde{X}\right) \\ &= \mathbb{E}[WY(1)\widetilde{X}] + \mathbb{E}[(1-W)Y(0)\widetilde{X}] = \pi_1\,\mathbb{E}[Y(1)\widetilde{X}] + \pi_0\,\mathbb{E}[Y(0)\widetilde{X}] \\ &= \pi_1\,\mathbb{C}\text{ov}\left(Y(1), X\right) + \pi_0\,\mathbb{C}\text{ov}\left(Y(0), X\right).\end{aligned} \tag{5.72}$$

Thus, when $\mathbb{C}\text{ov}\left(Y(0), X\right) = \mathbb{C}\text{ov}\left(Y(1), X\right)$, or when in the case of $\pi_0 = \pi_1$ (which can be obtained by complete randomisation), then $\xi = \xi_*$, so that

$$n\,\mathbb{V}\text{ar}\left(\widehat{\text{ATE}}_I\right) \xrightarrow{\mathbb{P}} \frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \frac{1}{\pi_0\pi_1}\xi_* V \xi_*^\top, \tag{5.73}$$

which, as the next theorem states, is the asymptotic variance of the ANCOVA II ATE estimator. Intuitively, one can make sense of this fact by realising that $\mathbb{C}\text{ov}\left(Y(0), X\right) = \mathbb{C}\text{ov}\left(Y(1), X\right)$ implies that the relation between the covariates and the outcome is the same across the treatment- and control groups. This means that no interaction effects between X and W exist, so that nothing is gained by using the ANCOVA II estimator instead of the ANCOVA I estimator. This fact is more formally derived in equation (5.80).

Theorem 5.3.4.
The ANCOVA II ATE estimator $\widehat{\text{ATE}}_{II}$ is consistent with $\sqrt{n}\left(\widehat{\text{ATE}}_{II} - \text{ATE}\right) \xrightarrow{\mathbb{P}} 0$ and has asymptotic variance given by

$$n\,\mathbb{V}\text{ar}\left(\widehat{\text{ATE}}_{II}\right) \longrightarrow \frac{\sigma_0^2}{\pi_0} - \frac{\sigma_1^2}{\pi_1} - \frac{1}{\pi_0\pi_1}\xi_* V \xi_*^\top, \tag{5.74}$$

for $n \to \infty$, where

$$\begin{aligned}\sigma_w^2 &= \mathbb{V}\text{ar}\left(Y(w)\right) \\ V &= \mathbb{V}\text{ar}(X)^{-1} \\ \xi_* &= \pi_0\,\mathbb{C}\text{ov}\left(Y(1), X\right) + \pi_1\,\mathbb{C}\text{ov}\left(Y(0), X\right).\end{aligned} \tag{5.75}$$

◀

Proof. The proof follows the structure of the proof of theorem 5.3.3, now with observations $D = (1, W, \widetilde{X}, W\widetilde{X})$. Following the same procedure, we show in appendix C.3 that the influence

function of the ANCOVA II ATE estimator, again dropping arguments of the functions, is given as $\varphi_{\mathrm{ATE},II} = \varphi_1 - \varphi_0$, where

$$\varphi_1 = \frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1)]\right) - \frac{W-\pi_1}{\pi_1}\widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(1)\right), \tag{5.76}$$

and

$$\varphi_0 = \frac{1-W}{\pi_0}\left(Y - \mathbb{E}[Y(0)]\right) - \frac{(1-W)-\pi_0}{\pi_0}\widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(0)\right), \tag{5.77}$$

so that, as we also derive in appendix C.3,

$$\varphi_{\mathrm{ATE},II} = \varphi_\Delta - \frac{W-\pi_1}{\pi_0\pi_1}\widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\xi_*^\top, \tag{5.78}$$

which is the same as $\varphi_{\mathrm{ATE},I}$ in equation (5.60) but with ξ_* in place of $\xi = \mathbb{C}\mathrm{ov}(X,Y)$, using the same definition of φ_Δ, defined in equation (5.61). The result now follows from the same steps as carried out in the proof of theorem 5.3.3. ■

Remark. In the above proof, Schuler et al. [35] derive that ξ^* is a common factor between φ_1 and φ_0, whereas we reach an expression with ξ^* only when we take the difference to obtain the influence function $\varphi_{\mathrm{ATE},II}$. Using ξ^* as a common factor, Schuler et al. appear to mistakenly use an assumption of constant treatment effect, that is

$$\mathrm{ATE} = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y(1)\,|\,X] - \mathbb{E}[Y(0)\,|\,X]. \tag{5.79}$$

In this case of constant treatment effect, we get that

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\left(Y(1), X\right) &= \mathbb{E}\left[\left(Y(1) - \mathbb{E}[Y(1)]\right)\left(X - \mathbb{E}[X]\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(Y(1) - \mathbb{E}[Y(1)]\right)\left(X - \mathbb{E}[X]\right)\,\middle|\,X\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(Y(1) - \mathbb{E}[Y(1)]\right)\,\middle|\,X\right]\left(X - \mathbb{E}[X]\right)\right] \\
&= \mathbb{E}\left[\left(\mathbb{E}\left[Y(1)\,\middle|\,X\right] - \mathbb{E}[Y(1)]\right)\left(X - \mathbb{E}[X]\right)\right] \\
&= \mathbb{E}\left[\left(\mathbb{E}\left[Y(0)\,\middle|\,X\right] - \mathbb{E}[Y(0)]\right)\left(X - \mathbb{E}[X]\right)\right] \\
&= \mathbb{C}\mathrm{ov}\left(Y(0), X\right),
\end{aligned} \tag{5.80}$$

where the assumption of constant treatment effect is used in the second to last equality, and the last equality follows from applying the previous equalities in reverse order. This results in a common factor in equations (5.76) and (5.77). This common factor is equal to ξ^*, since

$$\xi_* = \pi_0 \,\mathbb{C}\mathrm{ov}\left(Y(1), X\right) + \pi_1 \,\mathbb{C}\mathrm{ov}\left(Y(0), X\right) = \mathbb{C}\mathrm{ov}\left(Y(0), X\right) = \mathbb{C}\mathrm{ov}\left(Y(1), X\right) \tag{5.81}$$

in the case of constant treatment effect. ◀

We can now use the previous theorem to obtain the following corollary.

Corollary 5.3.5.
Adding covariates which are not a linear combination of the covariates in X to the ANCOVA II estimator cannot increase its asymptotic variance. ◀

Proof. Start by considering covariates X and an additional covariate $M \in \mathbb{R}$ which is not a linear combination of the covariates in X. We then want to consider the difference in asymptotic variance of the ATE estimator when using the ANCOVA II model with X contra (X, M). We will use the notation

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(X) &= \Sigma_X \\
\mathbb{V}\mathrm{ar}(M) &= \sigma_M^2 \\
\mathbb{C}\mathrm{ov}(M, X) &= \zeta \\
\xi_{X*} &= \pi_0 \,\mathbb{C}\mathrm{ov}\left(Y(1), X\right) + \pi_1 \,\mathbb{C}\mathrm{ov}\left(Y(0), X\right) \\
\xi_{M*} &= \pi_0 \,\mathbb{C}\mathrm{ov}\left(Y(1), M\right) + \pi_1 \,\mathbb{C}\mathrm{ov}\left(Y(0), M\right).
\end{aligned} \tag{5.82}$$

Using theorem 5.3.4 for the ANCOVA II model with covariates X, we obtain an asymptotic variance of $n \,\mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{II}\right)$ as

$$\frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \frac{1}{\pi_0 \pi_1} \xi_{X*} \Sigma_X^{-1} \xi_{X*}^\top. \tag{5.83}$$

When we further adjust by the covariate M we obtain by the same result that $n \,\mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{II}\right)$ has asymptotic variance

$$\begin{aligned}
&\frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \frac{1}{\pi_0 \pi_1} \left[\pi_0 \,\mathbb{C}\mathrm{ov}\left(Y(1), (X, M)\right) + \pi_1 \,\mathbb{C}\mathrm{ov}\left(Y(0), (X, M)\right)\right] \\
&\begin{bmatrix} \Sigma_X & \zeta^\top \\ \zeta & \sigma_M^2 \end{bmatrix}^{-1} \left[\pi_0 \,\mathbb{C}\mathrm{ov}\left(Y(1), (X, M)\right) + \pi_1 \,\mathbb{C}\mathrm{ov}\left(Y(0), (X, M)\right)\right]^\top \\
= &\frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \frac{1}{\pi_0 \pi_1} \begin{bmatrix} \xi_{X*} & \xi_{M*} \end{bmatrix} \begin{bmatrix} \Sigma_X & \zeta^\top \\ \zeta & \sigma_M^2 \end{bmatrix}^{-1} \begin{bmatrix} \xi_{X*}^\top \\ \xi_{M*} \end{bmatrix}.
\end{aligned} \tag{5.84}$$

In appendix C.4, we show that the difference between the asymptotic variance of the estimator using X and the asymptotic variance of the estimator using (X, M) can be expressed as

$$\frac{1}{\pi_0\pi_1}\frac{\left(\xi_{M*} - \xi_{X*}\Sigma_X^{-1}\zeta^\top\right)^2}{\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top}. \tag{5.85}$$

We now want to argue that this difference is non-negative, proving the result that the difference in asymptotic variance is non-negative. This is the case if the denominator $\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top$ is positive.

The determinant of $\mathbb{V}\mathrm{ar}\left((X, M)\right)$ is positive since the matrix is positive definite. Explicitly writing out the determinant with a useful formula for the determinant of a block matrix [61], we get

$$\det\left(\mathbb{V}\mathrm{ar}\left((X, M)\right)\right) = \det\left(\begin{bmatrix}\Sigma_X & \zeta^\top \\ \zeta & \sigma_M^2\end{bmatrix}\right) = \det\left(\Sigma_X^{-1}\right)\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right) > 0. \tag{5.86}$$

Since Σ_X^{-1} is positive definite, we have that $\det\left(\Sigma_X^{-1}\right) > 0$, so $\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top > 0$, which completes the proof. ■

Later, we will adjust for the estimated prognostic score in our ANCOVA II model. This result is relevant in this regard, since it shows that we can additionally include baseline covariates without increasing the asymptotic variance, when using the ANCOVA II estimator. Later, we will discuss the benefits of doing so in further detail. In the next theorem we will relate the asymptotic variances for the difference-in-means, ANCOVA I and ANCOVA II ATE estimators.

Theorem 5.3.6.
The ANCOVA II ATE estimator is more asymptotically efficient than the ANCOVA I estimator and difference-in-means estimator. There exists cases where the ANCOVA I ATE estimator is not more asymptotically efficient than the difference-in-means estimator for some $\pi_1 \neq \pi_0$ and $\mathbb{C}\mathrm{ov}\left(Y(0), X\right) \neq \mathbb{C}\mathrm{ov}\left(Y(1), X\right)$. In case $\pi_1 = \pi_0$ or $\mathbb{C}\mathrm{ov}\left(Y(0), X\right) = \mathbb{C}\mathrm{ov}\left(Y(1), X\right)$ the ANCOVA I estimator is as asymptotically efficient as the ANCOVA II estimator.

Specifically, abusing the notation $\mathbb{V}\mathrm{ar}$, here denoting the asymptotic variance multiplied by n, we have

$$\begin{aligned}
&\mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{II}\right) \leqslant \mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{I}\right) \\
&\mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{II}\right) \leqslant \mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{\Delta}\right) \\
&\mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{I}\right) > \mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{\Delta}\right) \text{ for some } \pi_0 \neq \pi_1 \text{ and } \mathbb{C}\mathrm{ov}\left(Y(0), X\right) \neq \mathbb{C}\mathrm{ov}\left(Y(1), X\right) \\
&\mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{I}\right) = \mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{II}\right) \text{ when } \pi_1 = \pi_0 \text{ or } \mathbb{C}\mathrm{ov}\left(Y(0), X\right) = \mathbb{C}\mathrm{ov}\left(Y(1), X\right).
\end{aligned} \tag{5.87}$$

◀

Proof. We first prove $\mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{II}\right) \leqslant \mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{I}\right)$ by subtracting the asymptotic variance in theorem 5.3.4 from the asymptotic variance in theorem 5.3.3, obtaining

$$\begin{aligned}
&\frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} + \frac{1}{\pi_0\pi_1}\xi V\xi^\top - 2\frac{1}{\pi_0\pi_1}\xi_* V\xi^\top - \left(\frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \frac{1}{\pi_0\pi_1}\xi_* V\xi_*^\top\right) \\
&= \frac{1}{\pi_0\pi_1}\left(\xi V\xi^\top - 2\xi_* V\xi^\top + \xi_* V\xi_*^\top\right) \\
&= \frac{1}{\pi_0\pi_1}\left(\xi V^{1/2}V^{1/2}\xi^\top - 2\xi_* V^{1/2}V^{1/2}\xi^\top + \xi_* V^{1/2}V^{1/2}\xi_*^\top\right) \\
&= \frac{1}{\pi_0\pi_1}\left(\xi V^{1/2} - \xi_* V^{1/2}\right)\left(V^{1/2}\xi^\top - V^{1/2}\xi_*^\top\right) \\
&= \frac{1}{\pi_0\pi_1}\left((\xi - \xi_*)V^{1/2}\right)\left(V^{1/2}(\xi^\top - \xi_*^\top)\right) \\
&= \frac{1}{\pi_0\pi_1}\left(V^{1/2}(\xi - \xi_*)^\top\right)^\top\left(V^{1/2}(\xi - \xi_*)^\top\right) \geqslant 0,
\end{aligned} \tag{5.88}$$

using that V is positive definite and hence has a well-defined square root $V^{1/2}$ as well as the fact that $\xi_* V^{1/2}V^{1/2}\xi^\top = \xi V^{1/2}V^{1/2}\xi_*^\top$ since it is a scalar.

To determine that $\mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{II}\right) \leqslant \mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{\Delta}\right)$ we subtract the asymptotic variance in theorem 5.3.4 from the asymptotic variance in lemma 5.3.2 and thereby obtain

$$\frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \left(\frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \frac{1}{\pi_0\pi_1}\xi_* V\xi_*^\top\right) = \frac{1}{\pi_0\pi_1}\xi_* V\xi_*^\top, \tag{5.89}$$

which is greater than 0, since V is a positive definite matrix.

We only need to find one case where $\mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{I}\right) > \mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{\Delta}\right)$ to prove that there exits cases where the ANCOVA I estimator is not more asymptotically efficient than the difference-in-means estimator. Using e.g. $\pi_1 = 5/6$, $\pi_0 = 1/6$, $\mathbb{C}\mathrm{ov}\left(Y(1), X\right) = 4$ and $\mathbb{C}\mathrm{ov}\left(Y(0), X\right) = 1$ with X being one dimensional, we subtract the asymptotic variance in theorem 5.3.3 from the asymptotic variance in lemma 5.3.2 and thereby obtain

$$\begin{aligned}
-\frac{1}{\pi_0\pi_1}\left(\xi V\xi^\top - 2\xi_* V\xi^\top\right) &= -\frac{36}{5}\left[\left(\frac{5}{6}\cdot 4 + \frac{1}{6}\right)^2\sigma_X^{-2} - 2\left(\frac{1}{6}\cdot 4 + \frac{5}{6}\right)\left(\frac{5}{6}\cdot 4 + \frac{1}{6}\right)\sigma_X^{-2}\right] \\
&= -\frac{36}{5}\sigma_X^{-2}\left(\left(\frac{21}{6}\right)^2 - 2\cdot\frac{9}{6}\cdot\frac{21}{6}\right) \leqslant 0.
\end{aligned} \tag{5.90}$$

This implies that the asymptotic variance of the ANCOVA I estimator in this case is greater than the asymptotic variance of the difference-in-means estimator.

The case of $\mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_I\right) = \mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{II}\right)$ when $\pi_1 = \pi_0$ or $\mathbb{C}\mathrm{ov}\left(Y(0), X\right) = \mathbb{C}\mathrm{ov}\left(Y(1), X\right)$ was discussed as a remark after the proof of theorem 5.3.3. Thus, the situation of $\mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_I\right) > \mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_\Delta\right)$ can only occur when both $\pi_1 \neq \pi_0$ and $\mathbb{C}\mathrm{ov}\left(Y(0), X\right) \neq \mathbb{C}\mathrm{ov}\left(Y(1), X\right)$, since if one of these assumptions is fulfilled, the second and fourth (in)equalities in (5.87) (which we just proved) ensure that $\widehat{\mathrm{ATE}}_I$ is asymptotically more efficient than $\widehat{\mathrm{ATE}}_\Delta$. ■

In the example providing proof of the third inequality in (5.87), the treatment and control groups are not equally sized, and the covariates and the outcome are not related in the same way across treatment- and control groups, which according to the argument at equation (5.80) implies a heterogeneous treatment effect. We see that in this case, it is more harmful in terms of asymptotic efficiency to adjust for covariates without adjusting for these interaction effects than to use the simple unadjusted difference-in-means estimator.

5.3.2 Oracle Estimators

Before considering the potential reduction in variance of ATE estimators when adjusting for estimated prognostic scores, we will consider adjustment for the true prognostic score. For this purpose, we will refer to an *oracle estimator* as an estimator with influence function being the efficient influence function. We recall from the remarks after equation (5.9) that such an estimator obtains the lowest asymptotic variance among all RAL estimators of the estimand. As we will show, in different scenarios, the oracle estimator corresponds to different (infeasible) ANCOVA ATE estimators using the true conditional mean $f(X) = \big(\mathbb{E}[Y(0) \,|\, X],\ \mathbb{E}[Y(1) \,|\, X]\big)$ or $f(X) = \mathbb{E}[Y(0) \,|\, X]$ in place of X. We begin by stating and proving a result which holds in the general case of a heterogeneous treatment effect.

Lemma 5.3.7.
Let $f(X)$ be an arbitrary (possibly multivariate) transformation of the covariates X. Then the ANCOVA II ATE estimator that uses $f(X) = \big(\mathbb{E}[Y(0) \,|\, X],\ \mathbb{E}[Y(1) \,|\, X]\big)$ in place of X has the lowest asymptotic variance among all RAL estimators with access to X. ◀

Proof. We wish to show that the influence function of the ANCOVA II ATE estimator using $f(X)$ in place of X coincides with the efficient influence function of the ATE. We begin by replacing X with $f(X)^\top \in \mathbb{R}^k$ in equation (5.78), obtaining

$$\begin{aligned}\varphi_{\mathrm{ATE},II,f(X)} = &\frac{W}{\pi_1}\big(Y - \mathbb{E}[Y(1)]\big) - \frac{1-W}{\pi_0}\big(Y - \mathbb{E}[Y(0)]\big) \\ &- \frac{W - \pi_1}{\pi_0\pi_1}\big(f(X) - \mathbb{E}[f(X)]\big)\,\mathbb{V}\mathrm{ar}\big(f(X)\big)^{-1}\xi_{f*}^\top,\end{aligned} \tag{5.91}$$

where $\xi_{f*} = \pi_0\,\mathbb{C}\mathrm{ov}\big(Y(1), f(X)\big) + \pi_1\,\mathbb{C}\mathrm{ov}\big(Y(0), f(X)\big)$. In appendix C.5 we show that

$\mathbb{V}\mathrm{ar}\left(f(X)\right)^{-1}\xi_{f*}^{\top} = (\pi_1, \pi_0)^{\top}$. Thus, we obtain

$$\begin{aligned}\varphi_{\mathrm{ATE},II,f(X)} &= \frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1)]\right) - \frac{1-W}{\pi_0}\left(Y - \mathbb{E}[Y(0)]\right) \\ &\quad - \frac{W-\pi_1}{\pi_0\pi_1}\begin{pmatrix}\mathbb{E}[Y(0)\,|\,X] - \mathbb{E}[Y(0)] \\ \mathbb{E}[Y(1)\,|\,X] - \mathbb{E}[Y(1)]\end{pmatrix}^{\top}\begin{pmatrix}\pi_1 \\ \pi_0\end{pmatrix} \\ &= \frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1)]\right) - \frac{1-W}{\pi_0}\left(Y - \mathbb{E}[Y(0)]\right) \\ &\quad - \frac{W-\pi_1}{\pi_0\pi_1}\Big(\pi_1\left(\mathbb{E}[Y(0)\,|\,X] - \mathbb{E}[Y(0)]\right) + \pi_0\left(\mathbb{E}[Y(1)\,|\,X] - \mathbb{E}[Y(1)]\right)\Big),\end{aligned} \tag{5.92}$$

which is the efficient influence function of the ATE as seen in corollary 5.2.4. This implies that using $f(X)$ in place of X in the ANCOVA II model yields the most efficient ATE estimator among all RAL estimators. ■

In the following two corollaries, we turn to the less general situation in which the treatment effect is homogeneous. As we see from equations (2.1) and (2.2), this corresponds to $\mathrm{ATE} = \mathrm{CATE}(X)$.

Corollary 5.3.8.
Assume that $\mathbb{E}[Y(1)\,|\,X] = \mathbb{E}[Y(0)\,|\,X] + \mathrm{ATE}$, that is, the treatment effect is constant across all values of the covariates X. Then the ANCOVA II ATE estimator with $f(X) = \mathbb{E}[Y(0)\,|\,X]$ in place of X has the lowest possible asymptotic variance among all RAL estimators with access to X. ◀

Proof. Similar to the proof of lemma 5.3.7, we wish to consider the influence function for the ANCOVA II ATE estimator, and we want to show that using $f(X) = \mathbb{E}[Y(0)\,|\,X]$ in place of X gives the efficient influence function of the ATE. In appendix C.6, we show that $\mathbb{V}\mathrm{ar}\left(f(X)\right)^{-1}\xi_{f*} = 1$. Using the influence function of the ANCOVA II ATE estimator in equation (5.78), this yields

$$\begin{aligned}\varphi_{\mathrm{ATE},II,f(X)} &= \frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1)]\right) - \frac{1-W}{\pi_0}\left(Y - \mathbb{E}[Y(0)]\right) \\ &\quad - \frac{W-\pi_1}{\pi_0\pi_1}\left(\mathbb{E}[Y(0)\,|\,X] - \mathbb{E}[Y(0)]\right),\end{aligned} \tag{5.93}$$

which is the efficient influence function under a constant treatment effect, since in this case the equation in corollary 5.2.4 reduces to equation (5.93), using

$$\begin{aligned}
&\pi_1 \left(\mathbb{E}[Y(0) \mid X] - \mathbb{E}[Y(0)]\right) + \pi_0 \left(\mathbb{E}[Y(1) \mid X] - \mathbb{E}[Y(1)]\right) \\
&= \pi_1 \left(\mathbb{E}[Y(0) \mid X] - \mathbb{E}[Y(0)]\right) + \pi_0 \left(\mathbb{E}[Y(0) \mid X] + \text{ATE} - \mathbb{E}[Y(1)]\right) \\
&= \pi_1 \left(\mathbb{E}[Y(0) \mid X] - \mathbb{E}[Y(0)]\right) + \pi_0 \left(\mathbb{E}[Y(0) \mid X] - \mathbb{E}[Y(0)]\right) \\
&= \mathbb{E}[Y(0) \mid X] - \mathbb{E}[Y(0)].
\end{aligned} \tag{5.94}$$

■

Corollary 5.3.9.
Assume a constant treatment effect. Then corollary 5.3.8 also holds for the ANCOVA I ATE estimator. ◀

Proof. Theorem 5.3.6 implies that the ANCOVA I and ANCOVA II ATE estimators have the same variance when $\mathbb{C}\text{ov}\left(Y(1), f(X)\right) = \mathbb{C}\text{ov}\left(Y(0), f(X)\right)$. By equation (5.80), this equality holds in the case of a constant treatment effect. The result now follows from corollary 5.3.8. ■

These results show that linear adjustment for the true prognostic score provides the best possible estimator of the ATE, in the sense that we obtain the lowest possible asymptotic variance. We will later use a prognostic model trained on data $(\mathbb{X}', \mathbb{Y}')$ to approximate $\mathbb{E}[Y' \mid X']$, and hence the "best case" is when the historical data has the same distribution as the trial control arm, since we in this case have $\mathbb{E}[Y' \mid X'] = \mathbb{E}[Y(0) \mid X]$.

Considering corollaries 5.3.8 and 5.3.9, we can conclude that when there is a constant treatment effect, the ANCOVA I and ANCOVA II ATE estimators adjusting for the true prognostic score (that is, prognostic scores using "perfectly estimated" outcomes of digital twins) both provide the most efficient estimate of the ATE among all RAL estimators. However, when the treatment effect is not constant, lemma 5.3.7 states that, using the ANCOVA II ATE estimator, it is necessary to additionally adjust for $\mathbb{E}[Y(1) \mid X]$ to obtain a similar result. In most situations, this is not feasible since it requires training a model to predict the outcome of patients in the treatment arm, which are often exposed to a novel treatment for which no training data exist.

Regarding whether to adjust for covariates or not, and how to adjust for them in that case, theorem 5.3.6 guarantees that using the ANCOVA II ATE estimator at least ensures the same or lower asymptotic variance than both the difference-in-means and ANCOVA I ATE estimators. Furthermore, theorem 5.3.6 states that using the ANCOVA I ATE estimator can potentially lead to a larger asymptotic variance than the difference-in-means ATE estimator, which is an argument in favor of always choosing the ANCOVA II ATE estimator with adjustment for the prognostic score. However, this might not be feasible in practice, according to the regulatory guidelines for covariate adjustment described in section 2.3.2, due to the reasoning done at the end of section 2.3.1.

In the next subsection, we will relate the results of oracle estimators to the more realistic setting where we seek to estimate the prognostic score instead of adjusting for the true prognostic score.

5.3.3 Adjustment using a Prognostic Model

Before being able to prove the properties of the ANCOVA ATE estimators using a prognostic model to adjust for the predicted outcome of digital twins, we first need to define some measure theoretical properties. Firstly, we define a sequence of random functions $\{f_n\}_{n\in\mathbb{N}}$ as *uniformly bounded* if there exists $K > 0$ such that $\mathbb{P}(|f_n(X)| \geqslant K) = 0$ for all $n \in \mathbb{N}$. Secondly, we say, for $k > 0$, that $\{f_n(X)\}_{n\in\mathbb{N}}$ *converges in* L^k towards the stochastic variable $f(X)$ if

$$\lim_{n\to\infty} \mathbb{E}_{X,f_n}\left[|f_n(X) - f(X)|^k\right] = 0, \tag{5.95}$$

denoted as $f_n(X) \xrightarrow{L^k} f(X)$. For convenience, we will sometimes refer to the sequence $\{f_n\}_{n\in\mathbb{N}}$ as just f_n.

Lemma 5.3.10.
Let $f\colon \mathcal{X} \to \mathbb{R}$ be a bounded function on a compact set $\mathcal{X}$, $f_n\colon \mathcal{X} \to \mathbb{R}$ be a sequence of uniformly bounded random functions such that $|f(X) - f_n(X)| \xrightarrow{L^2} 0$, where $X \in \mathcal{X}$ is a random variable independent of f_n. Assume furthermore that $\mathbb{V}\mathrm{ar}_X\left(f_n(X)\right) > \varepsilon$ for all $n \in \mathbb{N}$ and some fixed $\varepsilon > 0$. Then

$$|f(X) - f_n(X)B_n| \xrightarrow{L^2} 0 \tag{5.96}$$

for

$$B_n = \frac{\mathbb{C}\mathrm{ov}_X\left(f(X), f_n(X)\right)}{\mathbb{V}\mathrm{ar}_X\left(f_n(X)\right)}. \tag{5.97}$$

◀

Proof. See appendix C.7. ■

Theorem 5.3.11.
Assume that the covariates X have compact support, that the average treatment effect is constant, $\mathbb{E}[Y(1) \mid X] = \mathbb{E}[Y(0) \mid X] + \mathrm{ATE}$, and that the function f defined as $f(x) = \mathbb{E}[Y(0) \mid X = x]$ is bounded. Furthermore, let m be a uniformly bounded random function learned from the external data $(\mathbb{X}', \mathbb{Y}')_{n'}$, which is independent of the current data, and $\left|m(X) - \mathbb{E}[Y(0) \mid X]\right| \xrightarrow{L^2} 0$ as $n' \to \infty$.

Then, if the number of participants n in the current trial increases such that $n = \mathcal{O}(n')$, the ANCOVA II ATE estimator that uses $m(X)$ in place of X is consistent and has the lowest possible asymptotic variance among all RAL estimators with access to X. ◀

Before proving the theorem, we note that m itself is a model which is trained on the historical data $(\mathbb{X}', \mathbb{Y}')$, so that it can be regarded as random with respect to the distribution and number n' of historical data points. Thus, in the proof, we will regard m as a sequence of random functions $\{m_{n'}\}_{n' \in \mathbb{N}}$, without explicitly denoting m as $m_{n'}$.

Proof. Throughout this proof, we will use the ANCOVA II estimator with $m(X)$ in place of the covariates X, which we denote as $\widehat{\mathrm{ATE}}$. The only case where the convergence in L^2 of m to $\mathbb{E}[Y(0) \,|\, X]$ is true while m is constant, is when it is constantly equal to the true prognostic score $\mathbb{E}[Y(0) \,|\, X]$. We know from corollary 5.3.8 that this estimator obtains the lowest possible asymptotic variance among all RAL estimators with access to X. Thus, for the rest of the proof, we assume that $m(X)$ is not numerically constant, that is, $\mathbb{V}\mathrm{ar}_X\left(m(X)\right) > \varepsilon$ for some $\varepsilon > 0$.

We will denote by $\widehat{\mathrm{ATE}}^*$ the oracle estimator described in corollary 5.3.8, meaning that $\widehat{\mathrm{ATE}}^*$ has the lowest possible asymptotic variance among all RAL estimators with access to X. Denoting this optimal asymptotic variance by ν_*^2, we have from the result in equation (5.9), that $\sqrt{n}\left(\widehat{\mathrm{ATE}}^* - \mathrm{ATE}\right) \xrightarrow{d} \mathcal{N}\left(0, \nu_*^2\right)$ as the number of current trial data points $n \to \infty$. If we can show that

$$\sqrt{n}\left(\widehat{\mathrm{ATE}} - \widehat{\mathrm{ATE}}^*\right) \xrightarrow{\mathbb{P}} 0, \tag{5.98}$$

for $n \to \infty$, we can use Slutsky's theorem to obtain

$$\sqrt{n}\left(\widehat{\mathrm{ATE}} - \mathrm{ATE}\right) = \sqrt{n}\left(\widehat{\mathrm{ATE}} - \widehat{\mathrm{ATE}}^*\right) + \sqrt{n}\left(\widehat{\mathrm{ATE}}^* - \mathrm{ATE}\right) \xrightarrow{d} \mathcal{N}\left(0, \nu_*^2\right), \tag{5.99}$$

for $n \to \infty$. Since the ATE estimator then converges in distribution, the moments converge towards the moment parameters in the distribution [59, p. 18]. Thus, we obtain the limiting variance and by lemma 5.3.1 we now have consistency of the estimator. We thus need to show equation (5.98) in order to prove the theorem. Later in the proof, we will see that in order for this to hold, we need in addition that the number of historical data points $n' \to \infty$, as $\widehat{\mathrm{ATE}}$ depends on m, which converges to the true prognostic score as $n' \to \infty$. Hence, to obtain the above convergence, we need both n and n' to go to ∞, which is obtained for $n = \mathcal{O}(n')$ when $n \to \infty$. However, for the first part of this proof, we will only vary n, thus considering m as fixed, and hence, all moments are taken with respect to the distribution of current data.

For the remainder of the proof, we drop the subscript of the influence function for ease of notation. The influence function in equation (5.78) for the ANCOVA II ATE estimator can, under a constant treatment effect and with $\widetilde{m(X)} = m(X) - \mathbb{E}_X[m(X)]$ in place of X, be expressed as

$$\begin{aligned}\varphi = &\frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1)]\right) - \frac{1-W}{\pi_0}\left(Y - \mathbb{E}[Y(0)]\right) \\ &- \frac{W - \pi_1}{\pi_0 \pi_1}\left(m(X) - \mathbb{E}_X[m(X)]\right)\mathbb{V}\mathrm{ar}\left(m(X)\right)^{-1}\mathbb{C}\mathrm{ov}\left(m(X), Y(0)\right)\end{aligned}$$

$$= \frac{W}{\pi_1}\big(Y - \mathbb{E}[Y(1)]\big) - \frac{1-W}{\pi_0}\big(Y - \mathbb{E}[Y(0)]\big) \tag{5.100}$$
$$- \frac{W-\pi_1}{\pi_0\pi_1}\big(m(X) - \mathbb{E}_X[m(X)]\big)\,\mathbb{V}\mathrm{ar}\big(m(X)\big)^{-1}\,\mathbb{C}\mathrm{ov}\Big(m(X), \mathbb{E}\big[Y(0) \mid X\big]\Big),$$

where we have replaced ξ_* with $\mathbb{C}\mathrm{ov}\big(m(X), Y(0)\big)$ due to (5.81), and we have used in the second equality that

$$\begin{aligned}\mathbb{C}\mathrm{ov}\big(m(X), Y(0)\big) = \mathbb{E}\Big[\widetilde{m(X)Y(0)}\Big] &= \mathbb{E}\Big[\widetilde{m(X)}\,\mathbb{E}[Y(0) \mid X]\Big] \\ &= \mathbb{C}\mathrm{ov}\big(m(X), \mathbb{E}[Y(0) \mid X]\big).\end{aligned} \tag{5.101}$$

We now define an ATE estimator as the sample mean

$$\widetilde{\mathrm{ATE}} := \overline{\varphi + \mathrm{ATE}} \tag{5.102}$$

and define the oracle counterpart as

$$\widetilde{\mathrm{ATE}}^* := \overline{\varphi^* + \mathrm{ATE}}, \tag{5.103}$$

where φ^* is the efficient influence function in equation (5.93), which corresponds to substituting $m(X)$ by $\mathbb{E}[Y(0) \mid X]$ in the influence function derived in equation (5.100). From corollary 5.2.5 we see, using the law of large numbers, that these sample means are consistent estimators of the ATE. We note that these estimators require that we know the true ATE, making them useless in practice, but we define them solely for the purpose of proving the theorem.

We will now show that $\widehat{\mathrm{ATE}}$ and $\widetilde{\mathrm{ATE}}$ share the same influence function. The estimator $\widetilde{\mathrm{ATE}}$ is obtained using the empirical distribution function $\widehat{F}_n$ to obtain a sample mean $\Psi(\widehat{F}_n) = \overline{\varphi + \mathrm{ATE}}$ as an estimator for the population mean estimand $\mathbb{E}_F[\varphi + \mathrm{ATE}]$, as seen in equation (5.5). Now using the efficient influence function for the population mean estimand from lemma 5.2.1, the influence function of the estimator $\widetilde{\mathrm{ATE}}$ can be expressed as

$$\breve{\varphi}(X, W, Y) = \varphi(X, W, Y) + \mathrm{ATE} - \mathbb{E}_{\widehat{F}_n}[\varphi(X, W, Y) + \mathrm{ATE}] = \varphi(X, W, Y), \tag{5.104}$$

where we use in the last equality that by corollary 5.2.5, the influence function has mean 0. Thus, the influence function for $\widetilde{\mathrm{ATE}}$ is the same as for the $\widehat{\mathrm{ATE}}$ estimator. Equivalent arguments can be used to obtain $\breve{\varphi}^* = \varphi^*$.

By theorem 5.3.4, we have that $\sqrt{n}\left(\widehat{\mathrm{ATE}} - \mathrm{ATE}\right) \xrightarrow{\mathbb{P}} 0$. Furthermore, since $\widehat{\mathrm{ATE}}$ and $\widetilde{\mathrm{ATE}}$ share the same influence function, we have from similar arguments as the ones in the proof of theorem 5.3.4 that $\sqrt{n}\left(\widetilde{\mathrm{ATE}} - \mathrm{ATE}\right) \xrightarrow{\mathbb{P}} 0$. From linearity of the probability limit, we then have

$$\sqrt{n}\left(\widehat{\mathrm{ATE}} - \mathrm{ATE}\right) - \sqrt{n}\left(\widetilde{\mathrm{ATE}} - \mathrm{ATE}\right) = \sqrt{n}\left(\widehat{\mathrm{ATE}} - \widetilde{\mathrm{ATE}}\right) \xrightarrow{\mathbb{P}} 0. \tag{5.105}$$

Similarly, we have $\sqrt{n}\left(\widehat{\mathrm{ATE}}^* - \widetilde{\mathrm{ATE}}^*\right) \xrightarrow{\mathbb{P}} 0$. Therefore, again using the linearity of the probability limit, if $\sqrt{n}\left(\widetilde{\mathrm{ATE}} - \widetilde{\mathrm{ATE}}^*\right) \xrightarrow{\mathbb{P}} 0$, then we will have

$$\begin{aligned}
&\sqrt{n}\left(\widehat{\mathrm{ATE}} - \widehat{\mathrm{ATE}}^*\right) \\
&= \sqrt{n}\left(\widehat{\mathrm{ATE}} - \widetilde{\mathrm{ATE}}\right) + \sqrt{n}\left(\widetilde{\mathrm{ATE}} - \widetilde{\mathrm{ATE}}^*\right) + \sqrt{n}\left(\widetilde{\mathrm{ATE}}^* - \widehat{\mathrm{ATE}}^*\right) \xrightarrow{\mathbb{P}} 0
\end{aligned} \tag{5.106}$$

as desired.

In order to show the remaining convergence $\sqrt{n}\left(\widetilde{\mathrm{ATE}} - \widetilde{\mathrm{ATE}}^*\right) \xrightarrow{\mathbb{P}} 0$, we will first consider $\widetilde{\mathrm{ATE}} - \widetilde{\mathrm{ATE}}^* = \overline{\varphi_1 - \varphi_0 - \varphi_1^* + \varphi_0^*}$. To obtain the desired convergence, we will show $\overline{\varphi_w - \varphi_w^*} \xrightarrow{\mathbb{P}} 0$ for $w = 0, 1$. First, note that we can obtain an expression of φ_1 in this case by using $\widetilde{m(X)} = m(X) - \mathbb{E}_X[m(X)]$ in place of $\widetilde{X}$ in equation (5.76), and similarly obtain an expression of φ_1^* by using demeaned $\mathbb{E}[Y(0) \mid X]$ in place of $\widetilde{X}$. Specifically, we obtain

$$\begin{aligned}
\overline{\varphi_1 - \varphi_1^*} &= \frac{1}{n}\sum_{i=1}^{n} \frac{W_i - \pi_1}{\pi_1}\left(\Big(\mathbb{E}[Y(0) \mid X_i] - \mathbb{E}\Big[\mathbb{E}[Y(0) \mid X]\Big]\Big) \frac{\mathbb{C}\mathrm{ov}\left(\mathbb{E}[Y(0) \mid X], Y(1)\right)}{\mathbb{V}\mathrm{ar}\left(\mathbb{E}[Y(0) \mid X]\right)}\right. \\
&\quad \left. - \big(m(X_i) - \mathbb{E}_X[m(X)]\big) \frac{\mathbb{C}\mathrm{ov}\left(m(X), Y(1)\right)}{\mathbb{V}\mathrm{ar}_X\left(m(X)\right)}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{W_i - \pi_1}{\pi_1}\left(\Big(\mathbb{E}[Y(0) \mid X_i] - \mathbb{E}\Big[\mathbb{E}[Y(0) \mid X]\Big]\Big) \frac{\mathbb{C}\mathrm{ov}\left(\mathbb{E}[Y(0) \mid X], \mathbb{E}[Y(0) \mid X]\right)}{\mathbb{V}\mathrm{ar}\left(\mathbb{E}[Y(0) \mid X]\right)}\right. \\
&\quad \left. - \big(m(X_i) - \mathbb{E}_X[m(X)]\big) \frac{\mathbb{C}\mathrm{ov}\left(m(X), \mathbb{E}[Y(0) \mid X]\right)}{\mathbb{V}\mathrm{ar}_X\left(m(X)\right)}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{W_i - \pi_1}{\pi_1}\left(\mathbb{E}[Y(0) \mid X_i] - m(X_i)B\right) \\
&\quad - \frac{1}{n}\sum_{i=1}^{n} \frac{W_i - \pi_1}{\pi_1}\left(\mathbb{E}[Y(0)] - \mathbb{E}_X[m(X)]B\right),
\end{aligned} \tag{5.107}$$

having denoted $B = \frac{\mathbb{C}\mathrm{ov}\left(m(X), \mathbb{E}[Y(0) \mid X]\right)}{\mathbb{V}\mathrm{ar}_X\left(m(X)\right)}$. The second equality follows from the fact that for any transformation $g(X)$ and under constant treatment effect, $\mathbb{C}\mathrm{ov}(g(X), Y(0)) = \mathbb{C}\mathrm{ov}(g(X), Y(1))$, as derived in (5.80), so that

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\left(g(X), Y(1)\right) &= \mathbb{C}\mathrm{ov}\left(g(X), Y(0)\right) \\
&= \mathbb{E}\left[g(X)Y(0)\right] - \mathbb{E}\left[g(X)\right]\mathbb{E}\left[Y(0)\right] \\
&= \mathbb{E}\Big[\mathbb{E}\left[g(X)Y(0) \mid X\right]\Big] - \mathbb{E}\left[g(X)\right]\mathbb{E}\left[Y(0)\right]
\end{aligned} \tag{5.108}$$

$$
\begin{aligned}
&= \mathbb{E}\left[g(X)\,\mathbb{E}\left[Y(0) \mid X\right]\right] - \mathbb{E}\left[g(X)\right]\mathbb{E}\left[\mathbb{E}\left[Y(0) \mid X\right]\right] \\
&= \mathbb{C}\mathrm{ov}\left(g(X), \mathbb{E}\left[Y(0) \mid X\right]\right).
\end{aligned}
$$

Similarly, using equation (5.77), we get that

$$
\begin{aligned}
\overline{\varphi_0 - \varphi_0^*} = &\frac{1}{n}\sum_{i=1}^{n} \frac{(1-W_i)-\pi_0}{\pi_0}\left(\mathbb{E}[Y(0) \mid X_i] - m(X_i)B\right) \\
&- \frac{1}{n}\sum_{i=1}^{n} \frac{(1-W_i)-\pi_0}{\pi_0}\left(\mathbb{E}[Y(0)] - \mathbb{E}_X[m(X)]B\right).
\end{aligned} \tag{5.109}
$$

We now show that both of these terms converge to 0 in L^2 with convergence rate $\sqrt{n}$ so that they both converge in probability to 0 with the same rate. We will consider the first term in each of equations (5.107) and (5.109), starting with the term $\frac{1}{n}\sum_{i=1}^{n} \frac{W_i-\pi_1}{\pi_1}\left(\mathbb{E}[Y(0) \mid X_i] - m(X_i)B\right)$ in equation (5.107). For this term, we must show that

$$
\mathbb{E}\left[\left(\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \frac{\widetilde{W}_i}{\pi_1}\left(\mathbb{E}[Y(0) \mid X_i] - m(X_i)B\right)\right)^2\right] \longrightarrow 0 \quad \text{for } n, n' \to \infty, \tag{5.110}
$$

with the expected value being with respect to the joint distribution of current and historical data, where the historical data only has an influence on the model m and consequently B, while $\widetilde{W}_i$, X_i and $Y_i(0)$ have distributions specified by the current data distribution. Thus, now we are in the setting of considering m and consequently B as stochastic through the historical data. Using the law of total expectation, conditioning on the historical data $(\mathbb{X}', \mathbb{Y}')$, we can rewrite this expression as

$$
\begin{aligned}
&\mathbb{E}\left[n\,\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\widetilde{W}_i}{\pi_1}\left(\mathbb{E}[Y(0) \mid X_i] - m(X_i)B\right)\right)^2 \;\middle|\; \mathbb{X}', \mathbb{Y}'\right]\right] \\
&= \mathbb{E}\left[n\,\mathbb{V}\mathrm{ar}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\widetilde{W}_i}{\pi_1}\left(\mathbb{E}[Y(0) \mid X_i] - m(X_i)B\right) \;\middle|\; \mathbb{X}', \mathbb{Y}'\right)\right] \\
&= \mathbb{E}\left[\mathbb{V}\mathrm{ar}\left(\frac{\widetilde{W}}{\pi_1}\left(\mathbb{E}[Y(0) \mid X] - m(X)B\right) \;\middle|\; \mathbb{X}', \mathbb{Y}'\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{\widetilde{W}}{\pi_1}\right)^2\right]\mathbb{E}\left[\left(\mathbb{E}[Y(0) \mid X] - m(X)B\right)^2 \;\middle|\; \mathbb{X}', \mathbb{Y}'\right]\right]
\end{aligned} \tag{5.111}
$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(\frac{\widetilde{W}}{\pi_1}\right)^2\right]\mathbb{E}\left[\left(\mathbb{E}[Y(0)\,|\,X] - m(X)B\right)^2\right] \\
&= \frac{1-\pi_1}{\pi_1}\,\mathbb{E}\left[\left(\mathbb{E}[Y(0)\,|\,X] - m(X)B\right)^2\right]
\end{aligned}
$$

In the first equality, we use that $\widetilde{W}_i$ is independent of X_i by randomisation and thus independent of $\mathbb{E}[Y(0)\,|\,X_i]$. Furthermore, $\widetilde{W}_i$ is independent of $m(X_i)$ and B by randomisation and the independence between the current and historical data. Thus, using that $\mathbb{E}[\widetilde{W}_i] = 0$ for all $i = 1, 2, \ldots, n$, the expected value of the sum of terms is 0, and we can use that $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ for any stochastic variable X.

In the second equality, we use that the model m is random in the sense that it depends on the historical data $(\mathbb{X}', \mathbb{Y}')$ but not on the current trial data $(\mathbb{X}, \mathbb{W}, \mathbb{Y})$, since the distributions of these data sets are assumed to be independent and m is only trained on the historical data. B is composed of (co)variances that are deterministic with respect to the current trial data. However, it is stochastic through m. Thus, conditioning on $(\mathbb{X}', \mathbb{Y}')$ in the inner expectation, we can regard B and m as being fixed, such that $m(X)$ is only stochastic through X in the current trial data. Since the current trial data is IID, the terms in the sum are then IID as well, which is why we can drop the subscript to denote the variance of a generic patient. We can thus write the variance of the sum as n times the variance of one of the terms and taking out $1/n^2$ from the sum, the n's cancel out.

In the third equality, we have re-written the variance as an expectation of the squared expression, and used that $\widetilde{W}$ is independent of both $\mathbb{E}[Y(0)\,|\,X]$ and B and by randomisation it is independent of $m(X)$. Furthermore, we have used that $\widetilde{W}$ does not depend on the historical data. In the fourth equality we use the law of total expectation. In the last equality we use

$$
\mathbb{E}\left[\left(\frac{\widetilde{W}}{\pi_1}\right)^2\right] = \mathbb{E}\left[\frac{(W-\pi_1)^2}{\pi_1^2}\right] = \mathbb{E}\left[\frac{W^2 + \pi_1^2 - 2W\pi_1}{\pi_1^2}\right] = \frac{\pi_1 + \pi_1^2 - 2\pi_1^2}{\pi_1^2} = \frac{1-\pi_1}{\pi_1}.
$$

Now considering the second term $\frac{1}{n}\sum_{i=1}^n \frac{W_i - \pi_1}{\pi_1}\left(\mathbb{E}[Y(0)] - \mathbb{E}_X[m(X)]B\right)$ in equation (5.107) we can use similar arguments as before to obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left(\sqrt{n}\frac{1}{n}\sum_{i=1}^n \frac{W_i - \pi_1}{\pi_1}\left(\mathbb{E}[Y(0)] - \mathbb{E}_X[m(X)]B\right)\right)^2\right] \\
&= \mathbb{E}\left[\mathbb{V}\mathrm{ar}\left(\frac{\widetilde{W}}{\pi_1}\left(\mathbb{E}[Y(0)] - \mathbb{E}_X[m(X)]B\right) \;\middle|\; \mathbb{X}', \mathbb{Y}'\right)\right] \\
&= \frac{1-\pi_1}{\pi_1}\,\mathbb{E}\left[\left(\mathbb{E}[Y(0)] - \mathbb{E}_X[m(X)]B\right)^2\right].
\end{aligned}
\tag{5.112}
$$

Considering the first term $\frac{1}{n}\sum_{i=1}^{n}\frac{(1-W_i)-\pi_0}{\pi_0}\left(\mathbb{E}[Y(0)\,|\,X_i]-m(X_i)B\right)$ in equation (5.109), we use equivalent arguments to obtain the expression

$$\begin{aligned}&\mathbb{E}\left[\left(\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{1-\widetilde{W}_i}{\pi_i}\left(\mathbb{E}[Y(0)\,|\,X_i]-m(X_i)B\right)\right)^2\right]\\&=\frac{1-\pi_0}{\pi_0}\,\mathbb{E}\left[\left(\mathbb{E}[Y(0)\,|\,X]-m(X)B\right)^2\right],\end{aligned}\tag{5.113}$$

and for the second term we get

$$\begin{aligned}&\mathbb{E}\left[\left(\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{1-\widetilde{W}_i}{\pi_1}\left(\mathbb{E}[Y(0)]-\mathbb{E}_X[m(X)]B\right)\right)^2\right]\\&=\frac{1-\pi_0}{\pi_0}\,\mathbb{E}\left[\left(\mathbb{E}[Y(0)]-\mathbb{E}_X[m(X)]B\right)^2\right].\end{aligned}\tag{5.114}$$

Now using the assumptions of (uniform) boundedness of m and the true prognostic score together with the assumption $\left|m(X)-\mathbb{E}[Y(0)\mid X]\right| \xrightarrow{L^2} 0$ for $n'\to\infty$, we can use lemma 5.3.10. Specifically, we can let $f_n(X)=m_{n'}(X)$ (denoted just as $m(X)$) and $f(X)=\mathbb{E}[Y(0)\mid X]$, in order to obtain

$$\left|m(X)B-\mathbb{E}[Y(0)\,|\,X]\right| \xrightarrow{L^2} 0 \quad \text{for} \quad n'\to\infty.\tag{5.115}$$

This implies that expressions in equations (5.111) and (5.113) converge to 0. Note here that the assumption of lemma 5.3.10 that m is independent of the current data X is ensured by m being trained only on independent historical data. Using equation (C.52) where we note that the convergence implies convergence in L^2, we have that

$$\left|\mathbb{E}_X\left[m(X)\right]-\mathbb{E}\left[\mathbb{E}[Y(0)\mid X]\right]\right| = \left|\mathbb{E}_X[m(X)]-\mathbb{E}[Y(0)]\right| \xrightarrow{L^2} 0 \quad \text{for} \quad n'\to\infty.\tag{5.116}$$

We can use this with lemma 5.3.10 to get

$$\left|\mathbb{E}_X[m(X)]B-\mathbb{E}[Y(0)]\right| \xrightarrow{L^2} 0 \quad \text{for} \quad n'\to\infty,\tag{5.117}$$

which implies that the expressions in equations 5.112 and (5.114) converge to 0 in L^2. We now have that both terms in equations (5.107) and (5.109) converge to 0 in L^2 as $n, n'\to\infty$. Thus, we have $\sqrt{n}\left(\widetilde{\mathrm{ATE}}-\widetilde{\mathrm{ATE}}^*\right)\xrightarrow{\mathbb{P}} 0$, giving that $\sqrt{n}\left(\widehat{\mathrm{ATE}}-\widehat{\mathrm{ATE}}^*\right)\xrightarrow{\mathbb{P}} 0$, hence obtaining the convergence in equation (5.99), proving the result. ■

Corollary 5.3.12.
Theorem 5.3.11 also holds for the ANCOVA I estimator. ◀

Proof. In the case of constant treatment effect, the ANCOVA I and II ATE estimators have the same asymptotic variance by corollary 5.3.9. ■

Theorem 5.3.11 and corollary 5.3.12 both hold in the case of a constant treatment effect. In this case, there is no need to use the ANCOVA II ATE estimator instead of its ANCOVA I counterpart. However, when a heterogeneous treatment effect is present, the ANCOVA II estimator is more efficient than the ANCOVA I estimator, as stated in theorem 5.3.6. In this case, the ANCOVA I estimator might even be less efficient than the difference-in-means estimator, which is an argument in favor of always using the ANCOVA II estimator instead of the ANCOVA I estimator, despite the result in corollary 5.3.12.

Theorem 5.3.11 ensures that under the assumptions of the theorem, when adjusting for an estimated prognostic score instead of e.g. baseline covariates, we will asymptotically get the most efficient estimator of the ATE. However, in light of corollary 5.3.5, we can additionally adjust for prognostic covariates; asymptotically, we cannot increase the variance of the ATE estimator by doing so. This is important in regard to regulatory limitations, which requires adjustment for e.g. the baseline value of the outcome variable, as described in section 2.3.2. However, as also described in this section, regulatory guidelines might prohibit the use of the ANCOVA II estimator, which implies a risk of less efficiency in the case of heterogeneous treatment effects, as described in the last paragraph.

Additionally, when having a finite sample, we could potentially decrease the variance of the ATE estimator by adjusting for strongly prognostic covariates directly in the ANCOVA II model in addition to the estimated prognostic score. Such a variance reduction may also be possible since if the treatment effect is heterogeneous, the ANCOVA II estimator with $m(X) = \widehat{\mathbb{E}}[Y(0) \mid X]$ in place of X is in general not the asymptotically most efficient estimator among all RAL estimators. That is, benefits in terms of variance reduction are in fact achievable, even asymptotically, when adjusting for baseline covariates in addition to the estimated prognostic score. However, it should be noted that when we include prognostic covariates directly in the ANCOVA model as well as in the prognostic model, we can not interpret the estimated parameter for this covariate. Furthermore, as noted earlier, adding covariates that are not correlated to the outcome causes a loss in degrees of freedom, potentially increasing the variance to some amount, or, if there seems to be correlation due to chance, we run the risk of overfitting and underestimating the variance, leading to loss of control over the type I error. However, due to the regulatory guidelines, adjustment by such variables should rarely happen.

The efficiency gain in corollary 5.3.5 depends on the correlation between the added covariates and the outcome, as seen by the difference in asymptotic variances in equation (5.85). Thus, prognostic models that are highly correlated with the outcome will give a high efficiency gain, but any presence of correlation could still decrease the variance. This also justifies the use of prognostic covariate adjustment for surrogate outcomes, that is when Y' and Y represent different but correlated outcomes. Furthermore, the prognostic model should be a nonlinear function of the covariates X, since in the linear case there would be no benefit beyond just including these in the ANCOVA model as raw covariate adjustments. This is the case since for any linear prognostic

model $M = m(X) = Xa$ for any vector $a \in \mathbb{R}^p$, the factor in the numerator in (5.85) reduces to

$$\begin{aligned}
\xi_{M*} - \xi_{X*}\Sigma_X^{-1}\zeta^\top &= \pi_0 \,\mathbb{C}\mathrm{ov}\left(Y(1), Xa\right) + \pi_1 \,\mathbb{C}\mathrm{ov}\left(Y(0), Xa\right) \\
&\quad - \Big(\pi_0 \,\mathbb{C}\mathrm{ov}\left(Y(1), X\right) + \pi_1 \,\mathbb{C}\mathrm{ov}\left(Y(0), X\right)\Big) \Sigma_X^{-1} \,\mathbb{C}\mathrm{ov}\left(Xa, X\right)^\top \\
&= \Big(\pi_0 \,\mathbb{C}\mathrm{ov}\left(Y(1), X\right) + \pi_1 \,\mathbb{C}\mathrm{ov}\left(Y(0), X\right)\Big) a \\
&\quad - \Big(\pi_0 \,\mathbb{C}\mathrm{ov}\left(Y(1), X\right) + \pi_1 \,\mathbb{C}\mathrm{ov}\left(Y(0), X\right)\Big) \Sigma_X^{-1} \,\mathbb{C}\mathrm{ov}\left(X, X\right) a \\
&= 0.
\end{aligned} \tag{5.118}$$

Thus using a linear prognostic model results in no efficiency gain. However, we have showed that the efficiency gain in (5.85) is always non-negative, so any gain must depend on the degree to which the prognostic model m is able to model correlation that is not contained in the direct linear correlation between the raw covariates X and the outcome Y. We note that in the case of choosing $m(X) = Xa$, the columns with $\mathbb{X}$ and $m(\mathbb{X})$ in the design matrix of the ANCOVA model become linearly dependent, defining an overspecified model in which there are no unique solutions for the parameters.

On the other hand, if we define $M = m(X) = Xa + g(X)$ for g being some non-linear function of X, we instead obtain that the factor in the numerator in (5.85) is

$$\begin{aligned}
&\xi_{M*} - \xi_{X*}\Sigma_X^{-1}\zeta^\top \\
&= \pi_0 \,\mathbb{C}\mathrm{ov}\left(Y(1), Xa + g(X)\right) + \pi_1 \,\mathbb{C}\mathrm{ov}\left(Y(0), Xa + g(X)\right) \\
&\quad - \Big(\pi_0 \,\mathbb{C}\mathrm{ov}\left(Y(1), X\right) + \pi_1 \,\mathbb{C}\mathrm{ov}\left(Y(0), X\right)\Big) \Sigma_X^{-1} \,\mathbb{C}\mathrm{ov}\left(Xa + g(X), X\right)^\top \\
&= \Big(\pi_0 \,\mathbb{C}\mathrm{ov}\left(Y(1), X\right) + \pi_1 \,\mathbb{C}\mathrm{ov}\left(Y(0), X\right)\Big) a + \pi_0 \,\mathbb{C}\mathrm{ov}\left(Y(1), g(X)\right) + \pi_1 \,\mathbb{C}\mathrm{ov}\left(Y(0), g(X)\right) \\
&\quad - \Big(\pi_0 \,\mathbb{C}\mathrm{ov}\left(Y(1), X\right) + \pi_1 \,\mathbb{C}\mathrm{ov}\left(Y(0), X\right)\Big) \Sigma_X^{-1} \left(\Sigma_X a + \mathbb{C}\mathrm{ov}\left(g(X), X\right)^\top\right) \\
&= \pi_0 \,\mathbb{C}\mathrm{ov}\left(Y(1), g(X)\right) + \pi_1 \,\mathbb{C}\mathrm{ov}\left(Y(0), g(X)\right) \\
&\quad - \Big(\pi_0 \,\mathbb{C}\mathrm{ov}\left(Y(1), X\right) + \pi_1 \,\mathbb{C}\mathrm{ov}\left(Y(0), X\right)\Big) \Sigma_X^{-1} \,\mathbb{C}\mathrm{ov}\left(g(X), X\right)^\top \\
&:= \xi_{g*} - \xi_{X*}\Sigma_X^{-1}\zeta_g^\top,
\end{aligned} \tag{5.119}$$

and the denominator is

$$\begin{aligned}
&\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top \\
&= \mathbb{V}\mathrm{ar}\left(Xa + g(X)\right) - \mathbb{C}\mathrm{ov}\left(Xa + g(X), X\right) \Sigma_X^{-1} \,\mathbb{C}\mathrm{ov}\left(Xa + g(X), X\right)^\top
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{V}\mathrm{ar}(Xa) + \mathbb{V}\mathrm{ar}\left(g(X)\right) + 2\,\mathbb{C}\mathrm{ov}\left(Xa, g(X)\right) \\
&\quad - \left(a^\top \Sigma_X + \mathbb{C}\mathrm{ov}\left(g(X), X\right)\right) \Sigma_X^{-1} \left(\Sigma_X a + \mathbb{C}\mathrm{ov}\left(g(X), X\right)^\top\right) \qquad (5.120) \\
&= a^\top \Sigma_X a + \mathbb{V}\mathrm{ar}\left(g(X)\right) + 2\,\mathbb{C}\mathrm{ov}\left(g(X), X\right) a \\
&\quad - \left(a^\top \Sigma_X a + 2\,\mathbb{C}\mathrm{ov}\left(g(X), X\right) a + \mathbb{C}\mathrm{ov}\left(g(X), X\right) \Sigma_X^{-1}\, \mathbb{C}\mathrm{ov}\left(g(X), X\right)^\top\right) \\
&= \mathbb{V}\mathrm{ar}\left(g(X)\right) - \mathbb{C}\mathrm{ov}\left(g(X), X\right) \Sigma_X^{-1}\, \mathbb{C}\mathrm{ov}\left(g(X), X\right)^\top \\
&:= \sigma_g^2 - \zeta_g \Sigma_X^{-1} \zeta_g^\top,
\end{aligned}
$$

so it is seen that the efficiency gain depends only on the non-linear part of the prognostic model.

5.4 Sample Size Calculations using Digital Twins

In this section we will describe how to determine a required sample size when utilising digital twins in RCT analyses. The efficiency gain when using a prognostic model can be exploited in order to decrease the sample size required in a trial, still maintaining the same power. Specifically, we wish to conduct sample size calculations for the ANCOVA II model, in the situation where we adjust for the prognostic score and possibly other raw covariates as well as interactions between these and the treatment allocation covariate. This is the situation described in section 3.4.3, where we need an approximation of the non-centrality parameter, as described in equation (3.60). However, this estimator potentially requires estimation of a lot of parameters, which induces a lot of uncertainty in the estimation, requiring very comprehensive sensitivity analyses, as described in a more simple setup in the end of section 3.2.

Instead, in the following corollary we determine the asymptotic variance of the ANCOVA II ATE estimator where we use the prognostic model without including any raw covariates in order to obtain an upper bound in the sense described in the corollary. This means that fewer parameters need to be estimated.

Corollary 5.4.1.
The asymptotic variance of $n\,\mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_{II}\right)$ for the ANCOVA II ATE estimator, where X is replaced by a transformation m of the covariates along with possible additional adjustment for some raw baseline covariates, is bounded by

$$
\frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \pi_1\pi_0 \left(\frac{\rho_1\sigma_1}{\pi_1} + \frac{\rho_0\sigma_0}{\pi_0}\right)^2, \qquad (5.121)
$$

where $\sigma_w^2 = \mathbb{V}\mathrm{ar}\left(Y(w)\right)$ and $\rho_w = \frac{\mathbb{C}\mathrm{ov}\left(Y(w), m(X)\right)}{\sigma_w \sigma_m}$ with $\sigma_m^2 = \mathbb{V}\mathrm{ar}_X\left(m(X)\right)$. ◀

Proof. From theorem 5.3.4, the asymptotic variance of the ANCOVA II estimator using $m(X)$ in place of X is

$$\begin{aligned}
&\frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \frac{1}{\pi_1\pi_0}\left(\pi_0 \operatorname{\mathbb{C}ov}\left(Y(1), m(X)\right) + \pi_1 \operatorname{\mathbb{C}ov}\left(Y(0), m(X)\right)\right)^2 \sigma_m^{-2} \\
&= \frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \frac{1}{\pi_1\pi_0}\left(\pi_0 \frac{\operatorname{\mathbb{C}ov}\left(Y(1), m(X)\right)}{\sigma_m} + \pi_1 \frac{\operatorname{\mathbb{C}ov}\left(Y(0), m(X)\right)}{\sigma_m}\right)^2 \\
&= \frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \frac{1}{\pi_1\pi_0}\left(\pi_0\rho_1\sigma_1 + \pi_1\rho_0\sigma_0\right)^2 \\
&= \frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \pi_1\pi_0\left(\frac{\rho_1\sigma_1}{\pi_1} + \frac{\rho_0\sigma_0}{\pi_0}\right)^2.
\end{aligned} \tag{5.122}$$

From corollary 5.3.5, we now conclude that this gives an upper bound when we further include the raw covariates in the ANCOVA II estimator. ■

Note that this result also provides a bound of the asymptotic variance for the ANCOVA I estimator in case of a constant treatment effect.

From this corollary we see that in order to determine the asymptotic variance we only need to determine the marginal outcome variances, and the correlation between the model and outcome in each treatment arm. When additionally including the raw covariates, we are unlikely to obtain a substantial decrease in variance, since the raw covariates are unlikely to provide further efficiency gain since their effect on the outcome should to a large extent be included in the prognostic model, so that we do not obtain an overly conservative estimate of the required sample size using this asymptotic variance. This is also seen since the upper bound provided for the asymptotic variance is always lower than the asymptotic variance of the unadjusted difference-in-means model, as seen by lemma 5.3.2. Thus, we get a mildly conservative estimation of the required sample size by using the upper bound given in corollary 5.4.1, but we do not expect the true required sample size to be much smaller than what we estimate using the upper bound.

More precisely, we can perform sample size calculations by utilising the methods described in section 3.4.1, but where we instead estimate the variance of the ATE estimator by the upper bound in corollary 5.4.1. Determining the sample size to obtain a certain power requires us to determine the smallest n such that equation (3.52) is satisfied. The degrees of freedom for the t-distribution can be chosen as the sample size minus the number of parameters in the full model including both the estimated prognostic score and baseline covariates, ensuring a conservative estimation; the variance estimation is carried out as if only the estimated prognostic score is in the model, but we still adjust the degrees of freedom according to the adjustment done for the full model. The expression depends on the non-centrality parameter for the t-distribution, which is the expected value of the entity in in equation (3.46), but with the true variance $\mathbb{V}\mathrm{ar}(\widehat{\mathrm{ATE}})$ in place of $\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\beta}_1 - \widehat{\beta}_0)$. Taking expected, the numerator becomes the difference between the assumed effect size and the chosen margin, and thus we obtain that the non-centrality parameter

can be expressed as

$$\frac{\beta_1 - \beta_0 - \Delta_s}{\sqrt{\frac{1}{n}\left(\frac{\sigma_0^2}{\pi_0} + \frac{\sigma_1^2}{\pi_1} - \pi_1\pi_0\left(\frac{\rho_1\sigma_1}{\pi_1} + \frac{\rho_0\sigma_0}{\pi_0}\right)^2\right)}}, \tag{5.123}$$

using the upper bound from corollary 5.4.1 in place of the variance in the denominator, where we have divided the expression in equation (5.121) by the sample size n.

A method for obtaining the estimates of the marginal outcome variances and the correlation between the model and outcome in each treatment arm is to use prior data of a control arm with a population similar to the current trial population. For instance if we have access to an outcome vector $\mathbb{Y}'' = \left(Y_1'', Y_2'', \ldots, Y_{n''}''\right)$ of n'' subjects independent from the historical data patients, with corresponding estimated prognostic scores $\mathbb{M}'' = \left(m_1'', m_2'', \ldots, m_{n''}''\right)$, we could estimate the control arm marginal outcome variance as

$$\widehat{\sigma}_0^2 = \frac{1}{n'' - 1}\sum_{i=1}^{n''}\left(Y_i'' - \overline{Y}''\right)^2. \tag{5.124}$$

Similarly, the correlation between the estimated prognostic scores and the outcome in the control arm can be estimated as

$$\widehat{\rho}_0 = \frac{\sum_{i=1}^{n''}\left(Y_i'' - \overline{Y}''\right)\left(m_i'' - \overline{m}''\right)}{\sqrt{\sum_{i=1}^{n''}\left(Y_i'' - \overline{Y}''\right)^2\sum_{i=1}^{n''}\left(m_i'' - \overline{m}''\right)^2}}. \tag{5.125}$$

To make a sensitivity analysis we can increase $\widehat{\sigma}_0^2$ or decrease $\widehat{\rho}_0$ using inflation and deflation factors.

The corresponding values for the treatment arm is often infeasible to estimate from prior data, since this data is often unavailable. Therefore it is often assumed that $\sigma_0 = \sigma_1$ and $\rho_0 = \rho_1$. The latter holds when the effect of the treatment is constant across the population. To be slightly more conservative one could assume a slightly higher value of σ_1 and smaller value of ρ_1 relative to the control arm counterparts, again using deflation and inflation factors. When assuming that

$\sigma_0 = \sigma_1 = \sigma_Y$ and $\rho_0 = \rho_1 = \rho$, the non-centrality parameter in (5.123) reduces to

$$\begin{aligned}
&\frac{\beta_1 - \beta_0 - \Delta_s}{\sqrt{\frac{1}{n}\left(\frac{\sigma_Y^2}{\pi_0} + \frac{\sigma_Y^2}{\pi_1} - \pi_1\pi_0\left(\frac{\rho\sigma_Y}{\pi_1} + \frac{\rho\sigma_Y}{\pi_0}\right)^2\right)}} \\
&= \frac{\beta_1 - \beta_0 - \Delta_s}{\sqrt{\frac{1}{n}\left(\frac{\sigma_Y^2(\pi_0+\pi_1)}{\pi_0\pi_1} - \pi_1\pi_0\left(\frac{\rho\sigma_Y(\pi_0+\pi_1)}{\pi_0\pi_1}\right)^2\right)}} = \frac{\beta_1 - \beta_0 - \Delta_s}{\sqrt{\frac{1}{n}\left(\frac{\sigma_Y^2}{\pi_0\pi_1} - \frac{\rho^2\sigma_Y^2}{\pi_0\pi_1}\right)}} \\
&= \frac{\beta_1 - \beta_0 - \Delta_s}{\sqrt{\frac{\sigma_Y^2(1-\rho^2)}{n\pi_0\pi_1}}} = \frac{\beta_1 - \beta_0 - \Delta_s}{\sqrt{\sigma_Y^2(1-\rho^2)}}\sqrt{n\pi_0\pi_1} \approx \frac{\beta_1 - \beta_0 - \Delta_s}{\sigma_Y\sqrt{(1-\rho^2)}}\sqrt{\frac{n_1 n_0}{n}} \\
&= \frac{\beta_1 - \beta_0 - \Delta_s}{\sigma_Y\sqrt{(1-\rho^2)}}\sqrt{\frac{r}{(1+r)^2}n},
\end{aligned} \tag{5.126}$$

where we have the approximate equality is due to $\pi_w \approx n_w/n$, which is true asymptotically from the law of large numbers, and the last equality follows from equation (3.20). We recognise this expression as the non-centrality parameter presented in equation (3.49).

When having multiple secondary confirmatory outcomes, each with a desired power level and target effect size, then the sample size should be calculated as above for each of these outcomes, having adjusted the power and choosing the largest sample size as shown in example 3.6.3. Therefore, this requires multiple prognostic models, one for each outcome of interest or possibly a prognostic model that can predict all at once.

6 | Simulation Study – Comparison of Approaches

In chapters 4 and 5, we described two different approaches for leveraging historical data in order to increase power in a current RCT with both a control and a treatment arm present, namely the approach of synthetic control arms and the approach of digital twins. In this chapter, we will use simulations in order to compare these approaches in terms of how well they perform, also comparing the use of different estimators within the digital twin setup. We will especially focus on how they perform in terms of increasing the power in a current RCT and if the methods provide control of the type I error rate. In addition, we will perform prospective power calculations using only historical data in order to see how well the power is estimated before conducting an RCT.

For the SCA approach based on one-to-one propensity score matching, we wish to test whether the procedure described in section 4.1.2 succeeds in lowering the variance of the ATE estimate while maintaining a fixed type I error under different scenarios. For the DT approach, section 5.3 provides us with nice asymptotic properties of efficiency among models not using synthetic control group patients, using well-established methods that strictly control the type I error. However, these results are asymptotic, so we find it relevant to test how the approach performs when used in finite samples, and when some of the assumptions of the asymptotic results are violated. We wish to test to which extent different prognostic models are able to increase power more than what would be achieved by adjusting for the covariates linearly when estimating the ATE. In addition, we wish to put to the test whether the DT approach is able to increase the power as much as the SCA approach while benefitting from the analytic guarantee of a fixed type I error.

The specific methods used in the simulation study are presented in section 6.1, and the results are presented in sections 6.2–6.4.2. All analyses were carried out using `R` version 4.1.1, and all code can be found here.

6.1 Methods

In the following, we will describe the methods used to generate the results of the simulation study. We begin by specifying the distribution of the simulated RCT data on which we will use our proposed methods. In addition, we specify which AN(C)OVA models we use with our proposed approaches, as well as the specific models used to estimate the prognostic score for the DT approach and the propensity score for the SCA approach. We then specify how we estimate entities for assessing the performance of the approaches and specific methods. Lastly, we specify the methods we use to prospectively estimate the power of a current RCT.

6.1.1 Distribution of Simulated Data

In order to compare different AN(C)OVA based approaches to estimate the ATE and to carry out prospective power estimation, we tested the methods on simulated historical and current RCT data in different scenarios. For this purpose, we implemented a function in `R` for simulating historical and current RCT data and a function for estimating the ATE, standard error, power, type I error, coverage and root mean squared error. In each scenario, we simulated $1,000$ pairs of current and historical RCT data sets with the same distribution, with the number of simulated patients in each data set and number of covariates specified later for each scenario. When generating data from an RCT, we used the distribution

$$Y(W) \,|\, X \sim \mathcal{N}\left(aX\mathbb{1}_{p\times p}X^\top + bX\mathbb{1}_{p\times 1} + cX\mathbb{1}_{p\times 1}W + \beta_W \cdot W,\ \sigma_y^2\right), \tag{6.1}$$

with $\mathbb{1}$ denoting a matrix or a vector (depending on the subscript) with 1 in each entry, and where the covariates, arranged in row-vector form as $X = (X_1, X_2, \ldots, X_p)$, are generated from a multivariate normal distribution with mean 0 and a covariance structure given as a matrix with variances of 1 in the diagonal and correlation coefficients of 0.3 in each of the off-diagonals. Choosing to generate data from $\mathbb{E}[X] = 0$ is arbitrary, but it allows us to generate data from some specified true $\mathrm{ATE} = \beta_W$ in a simple way, following the same argument as by equation (5.38). The treatment assignment W is simulated in accordance with a deterministic allocation scheme, assigning the first n_0 patients the value $W = 0$ and the remaining n_1 patients $W = 1$. This ensures complete independence between the observations, since the W's, being deterministic, are independent. Thus, we have the balancing property derived in example 2.2.2 since all covariates are simulated from the same distribution, and we can expect the results from chapter 5 to hold while being able to choose a fixed allocation ratio.

We generate historical data using the same distribution as in (6.1), but where the last two terms are 0, since there are no treated subjects. We also allow the simulated historical data having a distributional shift d for the covariates, meaning that we simulate the covariates from a normal distribution with mean d instead of mean 0. Such a shift could occur if the current RCT has different inclusion and/or exclusion criteria than for the trials in the historical data. We note that in this case, the ATE in the historical data is different from the ATE in the current data when $c \neq 0$, but this does not have an influence on $Y(0) \,|\, X$, which is the outcome we seek to predict with the prognostic model in the DT approach. For the PSM approach, this makes the outcome of current control group patients and historical control groups comparable when adjusting for all confounders in the modelled propensity score.

The value of a controls the non-linearity and interaction effects between the covariates, while the values of b, c and d control the linear main effects, interaction effects between the covariates and treatment assignment, and distributional shift, respectively. With inspiration from Schuler et al. [35] we will consider four different scenarios under which data is generated, determined by the values of these parameters. In table 6.1, these scenarios and corresponding values of parameters are listed.

In all analyses, we specified the noise term as $\sigma_y^2 = 1$ and the true ATE as $\text{ATE} = 3$ with the intent to demonstrate superiority with a margin of $\Delta_s = 1$, specifying the null hypothesis as $\text{ATE} \leqslant \Delta_s$. All tests were performed using a significance level of 2.5% for the one-sided superiority tests.

Scenario	$\boldsymbol{a}$	$\boldsymbol{b}$	$\boldsymbol{c}$	$\boldsymbol{d}$
Linear covariate effects	0	1	0	0
Homogeneous treatment effect	0.5	1	0	0
Heterogeneous treatment effect	0.5	1	1	0
Covariates shifted	0.5	1	1	2

Table 6.1: Choices of coefficients for the data generating process in both RCT and historical data in each simulated scenario.

6.1.2 AN(C)OVA Models for ATE Estimation

In each of the four scenarios, we compared the DT and SCA approaches for leveraging historical data. In addition, we compared these to an ANOVA model with design matrix

$$\mathbb{D} = [1 \ \mathbb{W}], \tag{6.2}$$

an ANCOVA I model with design matrix

$$\mathbb{D} = [1 \ \ \mathbb{W} \ \ \mathbb{X}], \tag{6.3}$$

and its ANCOVA II model counterpart having design matrix

$$\mathbb{D} = \left[1 \ \ \mathbb{W} \ \ \widetilde{\mathbb{X}} \ \ \text{diag}_n(\mathbb{W})\widetilde{\mathbb{X}}\right], \tag{6.4}$$

neither of which leverage historical data.

For the SCA approach with PSM, we use these same AN(C)OVA estimators. For the DT approach, we construct the prognostic model m with methods that will be described in section 6.1.3. For all DT approaches, we use equivalent estimators, but where $m(\mathbb{X})$ is included in the design matrix. Specifically, we use the most simple ANCOVA I estimator with design matrix

$$\mathbb{D} = \left[1 \ \ \mathbb{W} \ \ m(\mathbb{X})\right], \tag{6.5}$$

as well as the ANCOVA I estimator additionally including raw covariate adjustment, having design matrix

$$\mathbb{D} = \left[1 \ \ \mathbb{W} \ \ m(\mathbb{X}) \ \ \mathbb{X}\right]. \tag{6.6}$$

Furthermore, we use their ANCOVA II counterparts, having design matrices

$$\mathbb{D} = \left[1 \;\; \mathbb{W} \;\; \widetilde{m}(\mathbb{X}) \;\; \mathrm{diag}_n(\mathbb{W})\widetilde{m}(\mathbb{X})\right], \tag{6.7}$$

and

$$\mathbb{D} = \left[1 \;\; \mathbb{W} \;\; \widetilde{\mathbb{X}} \;\; \widetilde{m}(\mathbb{X}) \;\; \mathrm{diag}_n(\mathbb{W})\widetilde{\mathbb{X}} \;\; \mathrm{diag}_n(\mathbb{W})\widetilde{m}(\mathbb{X})\right], \tag{6.8}$$

respectively.

In all AN(C)OVA analyses, we used the standard model dependent variance estimators of the ATE estimates given in (3.29) for ANOVA and (3.43) for ANCOVA. Using the performance measures in section 6.1.4, we are able to empirically investigate the claim by Wang et al. [36] that the model dependent variance estimators are in themselves robust to misspecification of the ANCOVA model, so that robust estimation as described in section 3.5 is not necessary.

6.1.3 Models for Prognostic Score and Propensity Score

In order to estimate the prognostic score for the DT approach, we used a range of different prognostic models, including the linear penalised LASSO model [62, pp. 241–248] and the non-linear random forest model [63, pp. 305–313, 587–597] [62, pp. 327–345], which, rather than imposing strict assumptions on the relation of the covariates and the outcome, is more data-driven.

In practice we are only able to adjust for a few covariates in the ANCOVA model, so even though nothing is gained from using an estimated prognostic score from a linear prognostic model over raw adjustment by the covariates, such a prognostic model enables us to exploit information from other covariates than the few adjusted for directly in the ANCOVA model. Thus, we wish to investigate if the LASSO model performs better than a standard linear model.

Our goal of using a random forest machine learning model for estimating the prognostic score in the DT approach is to illustrate the benefits of gaining larger precision of the prognostic model by modelling the more realistic situation of non-linearity and interaction effects.

In order to construct the prognostic model from these models, a number of tuning parameters need to be specified. For the LASSO model, we need to specify the penalty parameter λ which penalises the size of the ML estimates of the parameters making these go towards 0 as λ increases. This value is chosen by 10-fold cross validation among a grid of 100 values of λ equally spaced on a log-linear scale in the range $[\lambda_{\min}, \lambda_{\max}]$, where $\lambda_{\min} = 0.0001$ is chosen close to 0 and $\lambda_{\max}$ is the smallest value of λ such that all coefficients are estimated as 0.

Using the random forest method requires us to specify the minimal node size $n'_{\min}$, the number of covariates m that can be used in each split, and the number of trees in the forest B. It would be computationally cumbersome to determine these by k-fold cross-validation, since we would have to perform bootstrap for k training data sets for a prespecified number of combinations of

$n'_{\min}, m$ and B. As an alternative to this approach, we chose $m = \lfloor p/3 \rfloor$ and $n'_{\min} = 5$, which is in accordance with the default choices for regression purposes suggested by [62, pp. 343–345] and [63, pp. 589, 592]. In regard to choosing the value of B, we can choose it sufficiently large to ensure that the prediction error rate has settled down. This can be done since B is not related to the problem of overfitting, so the cost of choosing a large B is only computational time. We chose $B = 500$ trees for the random forest model.

Furthermore, we will use a simple linear model as the prognostic model in order to investigate whether non-linear models enable larger efficiency when data is generated non-linearly. In addition, we will use a prediction model which predicts the prognostic score as a uniformly random generated scalar in the range of outcomes in the current RCT control group patients. This is carried out in order to investigate how the DT approach performs in a situation where the prediction model has extremely bad performance.

In addition, we will benchmark the PSM and DT approaches against the most optimal estimator in the DT approach, namely the oracle ATE estimator presented in lemma 5.3.7, which is infeasible in practice since it requires knowledge of the true underlying data distribution. In order to benchmark the DT approach against a "perfect" digital twin, we will compare the models using estimated prognostic scores not only to the oracle estimator in lemma 5.3.7 but also the oracle estimator adjusting only for the true prognostic score $\mathbb{E}[Y(0) \mid X]$ without additionally adjusting for $\mathbb{E}[Y(1) \mid X]$ as for the oracle estimator. We will refer to this ATE estimator as the *oracle0 estimator*. The true prognostic scores used by the oracle and oracle0 ATE estimators are obtained from the true distribution of the simulated data. We note that by corollary 5.3.8, the oracle and oracle0 estimators are asymptotically equivalent in the case of a homogeneous treatment effect.

We note that by lemma 5.3.7 and corollary 5.3.5, the traditional oracle estimator, which has the design matrix (6.7), has the same asymptotic variance as the oracle estimator using the design matrix (6.8). However, due to the assumption of homogeneous treatment effect in corollary 5.3.8, in presence of a heterogeneous treatment effect, the traditional oracle0 estimator with design matrix (6.7) has a potentially larger asymptotic variance than the one obtained from using the design matrix (6.8). We therefore expect that in general, the DT approach estimators that uses the design matrix in (6.8) are the most efficient.

We could as well have explored whether the models used for estimating the prognostic score could as well be used for constructing the propensity score in order to enhance performance of the PSM method. However, we will restrict ourselves to estimating propensity scores using a simple logistic regression model, and we will match directly on the propensity score using greedy matching without replacement. Greedy matching, as noted at the start of section 4.1.1, performs as well as optimal matching, but is much less computationally cumbersome. We estimate the variance with equation (4.18) using $B = 100$, and $\widehat{u}_b$ as the standard normal linear model estimate of the variance without using robust methods. That is, we limit our focus to compare the DT approach with a somewhat standard PSM procedure adapted to the context of two-arm RCTs.

Our purpose lies not in fine-tuning parameters and specific model choices as well as training

procedures for gaining the most optimal models. Therefore, better results regarding increased power could possibly be obtained by trying out different models than the ones we propose, and by fine-tuning these models.

6.1.4 Performance Measures

For each of the methods, we estimated the ATE and calculated the t-test statistic based on the variance estimate from the individual AN(C)OVA models. In order to measure performance, we estimated the power, coverage, probability of type I error and the root mean squared error of the ATE estimate. Furthermore, in order to measure the performance of the prognostic model in the DT approach, we estimated the L^2 norm of the difference between the true prognostic score and the one estimated by the prognostic model.

Specifically, we estimated the coverage as the proportion of times the true ATE was inside the confidence intervals

$$\left[\widehat{\mathrm{ATE}}_i - t_{0.975,n-k}\sqrt{\mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_i\right)},\ \ \widehat{\mathrm{ATE}}_i + t_{0.975,n-k}\sqrt{\mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_i\right)}\right] \tag{6.9}$$

for $i = 1, 2, \ldots, 1000$, where $\widehat{\mathrm{ATE}}_i$ and $\mathbb{V}\mathrm{ar}\left(\widehat{\mathrm{ATE}}_i\right)$ are the ATE and variance estimates from the ith data set, and $t_{0.975,n-k}$ is the 97.5%-quantile of the t-distribution with degrees of freedom equal to the sample size n minus the number of columns in the design matrix k. Using the 97.5%-quantile, we specify a significance level of 2.5% for the one-sided superiority test that we want to perform, which corresponds to specification of a 5% significance level for a two-sided test, and we would therefore expect an estimated coverage of 95%.

In order to empirically estimate the probability of correctly rejecting this null hypothesis (the power) as well as the probability of mistakenly rejecting the null hypothesis (the type I error probability), we simulated the $1,000$ pairs of data sets setting $\mathrm{ATE} = 3$ (false null hypothesis). We then estimated the power as the proportion among the $1,000$ data sets in which we were able to (correctly) reject the null hypothesis from the t-test statistic used with degrees of freedom dependent on the number of parameters included in the model. In order to estimate the probability of a type I error, we altered the simulated data by subtracting $\mathrm{ATE} - \Delta_s = 2$ from the outcome variable for patients in the treatment group, such that the true $\mathrm{ATE} = 1$ was equal to the superiority margin (the case of correct null hypothesis which has largest probability of rejection). We then estimated the type I error probability by calculating the number of times we (incorrectly) rejected the null hypothesis from the t-test statistic.

We define the *root mean squared error* across the $1,000$ data sets as

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{1000} \frac{1}{1000}\left(\widehat{\mathrm{ATE}}_i - \mathrm{ATE}\right)^2}, \tag{6.10}$$

which is effectively a sample estimate of the standard deviation of $\widehat{\mathrm{ATE}}_i$ when the estimator is unbiased.

Lastly, we estimated the L^2 norm of the difference between the true and estimated prognostic score by the empirical mean

$$\widehat{\mathbb{E}}\left[\left|m(X) - \mathbb{E}\left[Y(0) \mid X\right]\right|^2\right] = \frac{1}{1000}\sum_{i=1}^{1000}\frac{1}{n}\sum_{j=1}^{n}\left|m_i(x_{ij}) - \mathbb{E}\left[Y(0) \mid X = x_{ij}\right]\right|^2, \quad (6.11)$$

where m_i is the prognostic model trained on the ith historical data set and x_{ij} is the vector of covariates belonging to the jth patient in the ith current trial data set. By evaluating the performance of the models on the current RCT data, we avoid the problem of using the same historical data for training and evaluating the models.

6.1.5 Prospective Power Estimation

For the ANOVA model not leveraging historical data, we prospectively estimated the power of the t-test before estimating the ATE by estimating the non-centrality parameter in equation (3.33) as $\widehat{c}(n)$ using the true $\mathrm{ATE} = 3$ from which data was simulated, the superiority margin $\Delta_s = 1$ and σ_Y estimated as the square root of the sample variance of the response variable in the historical data. Under the true alternative hypothesis, we then have approximately that $T_{\mathrm{ANOVA}} \sim t\left(n-2,\, \widehat{c}(n)\right)$, which we can use to estimate the power as

$$\widehat{1-\beta} = \widehat{\mathbb{P}}\left(T_{\mathrm{ANOVA}} > t_{0.975,n-2}\right) = 1 - F_{t,n-2,\,\widehat{c}(n)}(t_{0.975,n-2}), \quad (6.12)$$

for $t_{0.975,n-2}$ denoting the critical value of the t-test statistic using the significance level $\alpha = 0.025$ for a one-sided test, and $F_{t,n-2,\,\widehat{c}(n)}$ being the cumulative distribution function of the non-central t-distribution with $n-2$ degrees of freedom and non-centrality parameter $\widehat{c}(n)$. Additionally, we used the Guenther and Schouthen approximation, which states that choosing n as in equation (3.37) should give a power of at least $1-\beta$. A bit of algebra then shows that by (3.37), one can obtain

$$\widehat{1-\beta} = \Phi\left(\sqrt{\frac{r}{(1+r)^2}\frac{(\mathrm{ATE}-\Delta_s)^2}{\widehat{\sigma}^2}\left(n - \frac{z_{0.975}^2}{2}\right)} - z_{0.975}\right), \quad (6.13)$$

for Φ denoting the cumulative distribution function of the standard normal distribution, and $z_{0.975}$ is the 97.5%-quantile of the standard normal distribution.

For the DT approach, we followed the method outlined in section 5.4. Specifically, we approximated the non-centrality parameter in (5.123) as $\widehat{c}(n)$, optimistically assuming that $\sigma_0 = \sigma_1 = \sigma_Y$ and $\rho_0 = \rho_1 = \rho$. In addition to estimating σ_Y similar as for the ANOVA model, we estimate ρ as the sample correlation between predictions made by m and actual outcomes, using equation (3.45). As discussed in section 5.4, we need independent data to properly carry this out. A

cross validation approach could solve this problem, but we found it to be too computationally cumbersome. Instead, we simulated additional historical data having the same distribution as the existing historical data, with $n'' = 500$. On this independent data, we estimated ρ by the sample correlation formula in (5.125). By following the same steps as for the ANOVA model, we then estimated the power as

$$\widehat{1-\beta} = \widehat{\mathbb{P}}\left(T_{\text{ANCOVA}} > t_{0.975,n-k}\right) = 1 - F_{t,n-k,\widehat{c}(n)}(t_{0.975,n-k}), \tag{6.14}$$

using that approximately $T_{\text{ANCOVA}} \sim t\left(n-k,\, \widehat{c}(n)\right)$ under the true alternative hypothesis. Additionally, we used the Frison-Pocock approximation in (3.54) to obtain

$$\widehat{1-\beta} = \Phi\left(\sqrt{\frac{r}{(1+r)^2}\frac{(\text{ATE}-\Delta_s)^2}{\widehat{\sigma}^2(1-\widehat{\rho}^2)}n} - z_{0.975}\right) \tag{6.15}$$

and the Guenther-Schouten approximation in (3.55) to obtain

$$\widehat{1-\beta} = \Phi\left(\sqrt{\frac{r}{(1+r)^2}\frac{(\text{ATE}-\Delta_s)^2}{\widehat{\sigma}^2(1-\widehat{\rho}^2)}\left(n - \frac{z_{0.975}^2}{2}\right)} - z_{0.975}\right). \tag{6.16}$$

For the ANCOVA model not leveraging historical data and the PSM approach, we used the multivariate equivalents to the Frison-Pocock and Guenther-Schouten approximation formulas in (3.62) and (3.63), substituting $\widehat{\rho}^2$ in (6.15) and (6.16) with $\widehat{R}^2$ from equation (3.58), which was estimated using standard sample (co)variance formulas on historical data.

6.2 Performance in Different Scenarios

In this section we wish to compare the performance of the different approaches for leveraging historical data to estimate the ATE presented earlier, in the scenarios listed in table 6.1. We will compare the PSM and DT approaches, where the prognostic model for the latter approach is constructed using the methods presented in the previous section.

In each scenario presented in table 6.1, we simulated $1,000$ data sets, each with $p = 10$ covariates, $n = 500$ current RCT participants ($n_0 = 200$ in the control group and $n_1 = 300$ in the treatment group) and $n' = 5,000$ historical data patients. All adjustment of raw baseline covariates included all 10 covariates, and the prognostic and propensity score models were trained on all 10 covariates.

Before considering all scenarios, we begin by examining the empirical distributions of ATE estimates obtained from all of our different proposed ATE estimators under the heterogeneous treatment effect scenario. Doing so, we restrict ourselves to only use some of the proposed estimators in one of the scenarios. All four scenarios are considered in figure 6.2, having additional

results presented in table D.1 in appendix D.1. Figure 6.1 compares the AN(C)OVA estimators using design matrices (6.2) and (6.4) with the ATE estimator obtained from PSM with the design matrix in (6.4), as well as estimators using the DT approach with design matrix (6.8) in the heterogeneous treatment effect scenario.

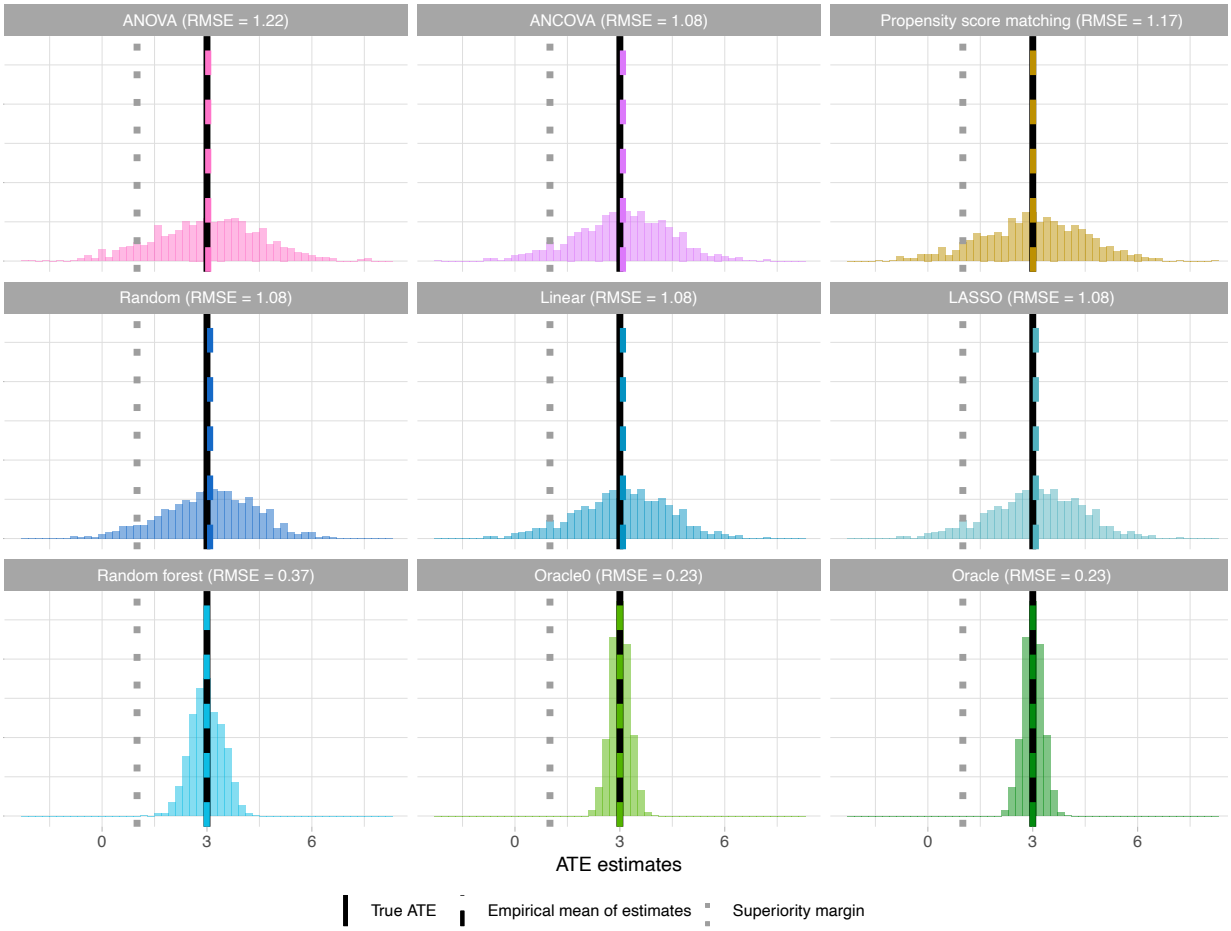


Figure 6.1: Empirical distributions of AN(C)OVA model ATE estimates obtained from $1,000$ simulated data sets under the heterogeneous treatment effect scenario. Additional results are found in table D.1 for some of the models.

From this figure, we see that the ANCOVA model achieves a better performance than the ANOVA model since it controls linearly for the covariates in the data generating process. Perhaps surprisingly, the PSM model has a worse performance than the ANCOVA model (but at least a better performance than the ANOVA model), even though the same raw covariates as in the ANCOVA model are adjusted for, and an additional 100 patients have been added to the control group based on their propensity scores. Even though we know of no analytical guarantee that the PSM method should provide a reduction in variance of the ATE estimate, we did not expect this result, and we suspect that the poor performance could be due to not specifying the propensity score model in accordance with the non-linear data generating process. Specifically, we have simulated the data and specified the prognostic model using the correct covariates such that the

condition in (4.3), and therefore also (4.4), hold. However, we suspect that the propensity score could more adequately ensure comparability between matched patients in regard to the outcome if the propensity score model was specified according to the non-linear relationships which data were generated from. This could be obtained from choosing the propensity score model as e.g. a random forest model.

Moving on to the ATE estimates obtained from the DT approach, we see firstly that additional adjustment for a badly estimated prognostic score (using the random prediction model) does not increase the standard error of the ATE estimate compared to the ANCOVA model ATE estimate by any amount that cannot be ascribed to random variation. Secondly, as expected from (5.118), no efficiency gain is obtained from using the linear prognostic model. This holds even though a LASSO penalty is introduced, presumably enhancing the predictive performance of the prognostic model. Again, this result is expected since the LASSO penalised model is a linear model. Lastly, we conclude that the random forest prognostic model provides a substantial decrease in the empirically estimated standard error of the ATE, which we expected to hold in some degree due to the random forest model being capable of modeling the non-linear and interaction effects included in the true underlying distribution of the simulated data.

The oracle0 estimator based on the true prognostic score should theoretically provide an asymptotic bound on the performance of the ATE estimator using any feasible model for the prognostic score. Indeed, we see that the oracle0 estimator has a better performance than the ATE estimator adjusting for the estimated prognostic score obtained from a random forest model. Even though data was simulated from a heterogeneous treatment effect, we detected no difference between the performances of the oracle0 and the asymptotically superior oracle estimator. From the results presented later in section 6.4, we suspect that this is due to the estimators converging relatively fast on our simulated data and that the ANCOVA models are specified with adjustments for a lot of covariates, diminishing the difference between the estimators due to possible overfitting.

In figure 6.2 we have summarised the empirical means of standard model dependent estimates of the standard error of the ATE estimates (filled points) and the empirically estimated RMSE (crosses) across the $1,000$ simulated data sets. This is done for all four scenarios presented in table 6.1 from each of the AN(C)OVA models with design matrices (6.2)–(6.8). Results of AN(C)OVA models not leveraging historical data as well as the propensity score method and the DT method using three of our proposed prognostic models, namely the random, random forest and the true prognostic score, are displayed. We choose to not include the linear, LASSO and oracle prognostic models, as we have just seen that (practically) nothing is to be gained compared to the ANCOVA and oracle0 model, respectively. In this way, we are able to benchmark the performance of a poor performing prognostic model, a prognostic model capable of modeling non-linear and interaction effects, and a "perfect" (infeasible) prognostic model in different scenarios. We note that for the PSM approach and the models which do not leverage historical data, the models with no raw covariate adjustment are the same between the "No interaction effects" and "Interaction effects" panels, both having design matrix (6.2). Table D.1 in appendix D.1 contains all exact results displayed on the figure. Specifically, along with empirical means of all estimates and estimated standard errors from each model, we report the RMSE and the

empirically estimated power, coverage and type I error rate.

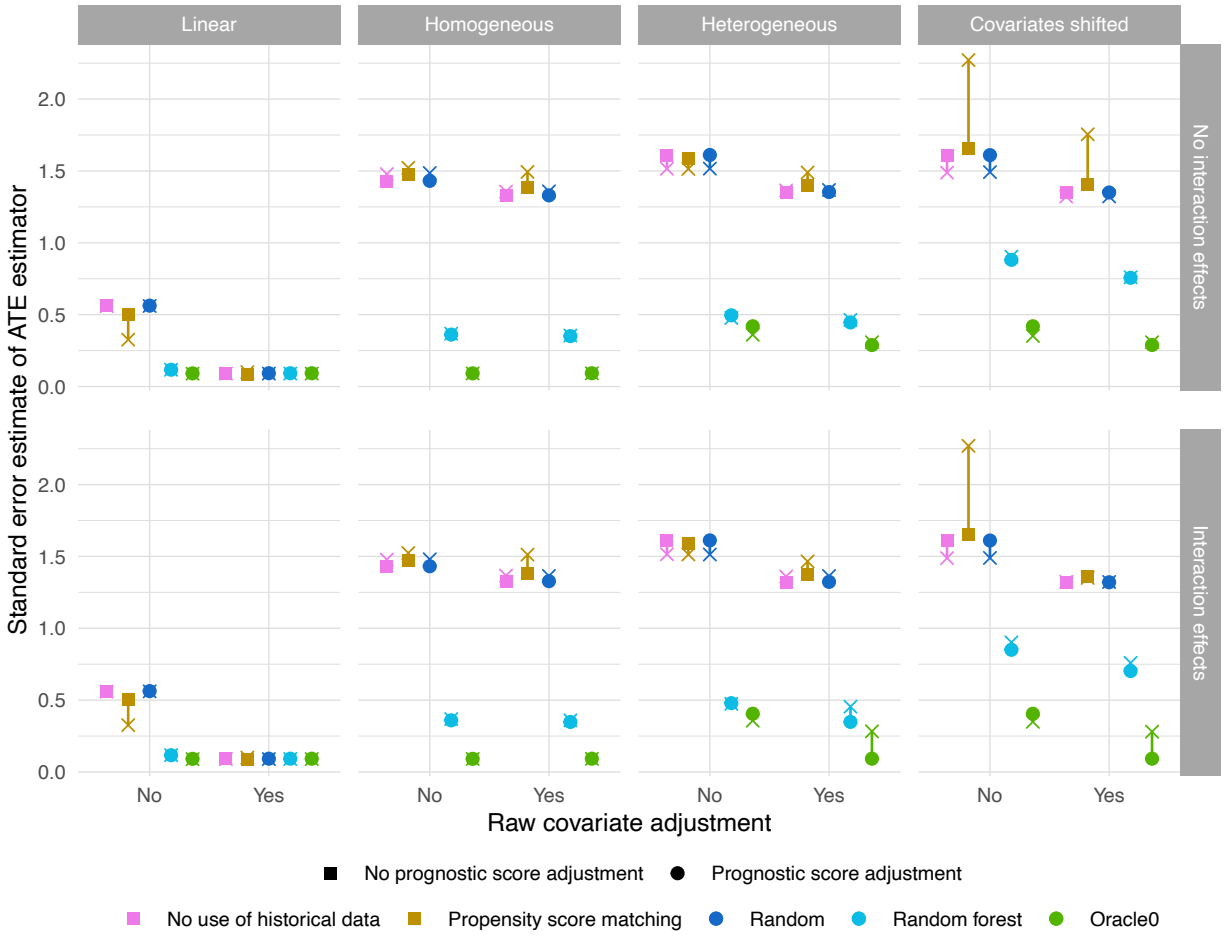


Figure 6.2: Standard error estimates (vertical axis) for AN(C)OVA model ATE estimates on $1,000$ simulated data sets under the four different scenarios (vertical panel classification). Filled points display empirical means of standard model dependent estimates across the $1,000$ simulated data sets. Crosses display the RMSE across the $1,000$ simulated data sets. Horizontal axis indicates whether raw covariate adjustments were included in the AN(C)OVA model for all 10 simulated covariates. Horizontal panel classification indicates whether interaction terms between treatment allocation and all raw covariates (as well as the estimated prognostic score for models "Random", "RF" and "Oracle0") was included in the AN(C)OVA model. Additional results are found in table D.1.

We note from figure 6.2 and table D.1 that the AN(C)OVA models adjusting for the prognostic score predicted by the random prognostic model have the same performance as the corresponding AN(C)OVA estimators not adjusting for this "prognostic score". Specifically, the method provides an ATE estimator with the same estimated standard error, control of type I error and proper coverage across all scenarios, indicating that the DT approach is robust to poor performing prognostic models. In general, the oracle0 estimator outperformed all other methods, followed by the random forest model, in terms of RMSE, average estimated standard error and thus power. The PSM approach provided only modest reductions in the estimated standard error for the linear co-

variate effects scenario, while even inflating the RMSE in most cases in the remaining scenarios. For a large proportion of the PSM estimators among all scenarios, the RMSE was larger than the standard model dependent standard error estimates, indicating that the method underestimates the standard error, leading to potential loss of control over the type I error. We again suspect that this could be due to the specification of the propensity score model, since the problem arises in the non-linear scenarios.

In the simplest scenario of linear covariate effects, the DT approach did not provide any efficiency gain compared to raw covariate adjustment. This was expected since the linear relationship is already modelled properly by the ANCOVA model, so that no additional non-linearities and interaction effects are left to be modeled by the prognostic model. In the case of no raw covariate adjustment, the PSM method provided a gain in terms of efficiency, but its performance was surpassed by its DT ANCOVA model counterparts using a random forest prognostic model. Moreover, comparing the RMSE with the average of the standard error estimates obtained from (4.18), the PSM approach seemed to overestimate the standard error when no raw covariate adjustments were conducted. This has the consequence that the method has the risk of being overly conservative in this specific scenario, meaning that the gain in terms of lower RMSE is not exploited completely in terms of power.

In the scenario of a homogeneous treatment effect, the DT approach using the random forest prognostic model provided a substantial decrease in the RMSE and estimated standard error, even compared to the ANCOVA model with adjustments for all covariates. That is, substantial benefits in terms of the standard error seems to be achievable by using a random forest prognostic model in the situation of non-linear relationships between the covariates and the potential outcome in the treatment group. Furthermore, all estimators based on the DT approach appear to give reasonable standard error estimates, which translates to an increase in power while delivering type I error control and proper coverage. On the other hand, when adjusting for raw covariates, the PSM approach seems to mildly underestimate the standard error, inflating the type I error rate to a minor degree. No models benefitted from including an interaction term since the underlying treatment effect is homogeneous.

In the heterogeneous treatment effect scenario, the tendencies are overall the same as for the homogeneous scenario. However, since a heterogeneous treatment effect is present, no analytical results from chapter 5 ensure asymptotic efficiency, but we see that a reasonable benefit in terms of efficiency can be achieved nonetheless. Perhaps surprisingly, the performance was not improved by including an interaction term, even though the underlying treatment effect is heterogeneous. This could be due to the only modest heterogeneity of the treatment effect obtained by a data generating process where the heterogeneity is specified by 10 terms, while the overall outcome is determined by 120 terms. Importantly, the type I error rate was highly inflated for the oracle0 estimator and to some degree for the estimator obtained from adjusting for the prognostic score estimated from a random forest model when both including all raw covariates and interaction effects, as seen in table D.1. From comparing the RMSE and the average estimated standard errors, we can see that this comes from a highly optimistic estimation of the standard error. However, the empirically estimated standard error suggests that the methods do provide the best

increase in power had we been able to properly estimate the standard error for the t-test statistic in each of the data sets. As described in section 3.4, analytic results regarding the ANCOVA model should ensure unbiasedness of the residual variance estimator when the model is correctly specified. However, in this case the model is near perfectly specified by including only the true prognostic score as a covariate, and the additional many raw covariate adjustments along with interaction terms thus specifies an excessively overspecified model. This leads to overfitting by the raw covariates randomly explaining variance in the noise and thus an overly optimistic estimate of the residual variance, making the estimated variance of the ATE estimate overly optimistic as well. The problem of inflated type I error rates was not solved by using heteroskedasticity-robust covariance matrix estimation for estimating the standard error of the ATE estimate. From this pitfall, we suggest using the digital twin approach with additional covariate adjustment for only one or a few raw baseline covariates, in accordance with the regulatory guidelines described in section 2.3.2.

For the DT approaches, the same tendencies hold for the scenario in which the covariate means of the historical data distribution are shifted by 2 compared to the distribution of the current RCT data. This indicates that overall, the DT approach is robust to the situation in which deviations between the training data of the prognostic model to some extent deviates from the current data distribution, when all relevant covariates are included. This could be due to the fact that even though the covariate distribution is shifted in the training data for the prognostic model, the relationships between the covariates and the outcome is still the same. Hence we can expect that if the prognostic model trained on the non-shifted historical data satisfies the condition of convergence in theorem 5.3.11, it should also hold for the shifted historical data, albeit with a slower convergence rate. We note however that the theorem relates to the case of a homogeneous treatment effect. Furthermore, if the prognostic model had been underspecified, the convergence would possibly not hold, due to omitted variable bias of the prognostic model.

For the PSM method in the case of the covariates being shifted, the large increase in RMSE points toward a severe problem, namely that the resulting ATE estimates seem to be biased, as can be seen in table D.1. Specifically, in the case of no raw covariate adjustment, we see from the table that the bias is negative, and consequently, the type I error rate is lower than the significance level. In the case of controlling for raw covariates, the bias is positive, and we get conversely that the type I error rate is above the significance level. For some reason, for our simulated data, the bias is negative when no raw covariate adjustments are conducted and positive when adjustments are carried out. The results are disappointing, considering that the situation of a covariate shift should be properly handled by the method described in section 4.1.2, even when not all confounding covariates are included. However, we again suspect that misspecification of the propensity score model could have an influence on the results.

6.3 Overspecification and Underspecification

So far, we have considered the situation where models are correctly specified in the sense that adjustment has only been done for the covariates in the data generating process. In this section we

wish to examine how the standard error of the ATE estimate is affected by the realistic scenario of overspecifying and underspecifying the prognostic model and the propensity score model, respectively. As we saw in examples 2.3.1 and 2.3.2, underspecifying linear models can lead to omitted variable bias, whereas overspecification only leads to inefficiency. Here we investigate not only linear models for the prognostic score, but we suspect that the same tendency is present for our proposed non-linear models. In section 6.2, we saw that as expected, under the most general conditions, the best performance among the DT approach estimators were achieved by using the non-linear random forest prognostic model. In this section, we will therefore limit our analyses of DT models using this prognostic model.

For the DT approach, underspecifying the prognostic model means, according to theorem 5.3.11, that we are not guaranteed to obtain an efficient ATE estimator, since the L^2 convergence does not hold in the case of a biased model. In addition, (5.85) shows that the efficiency gain of adjusting for the estimated prognostic score depends on the covariance between the estimate and the outcome, which is smaller for an inefficient model obtained from overspecification. However, a model like the random forest model should itself conduct proper variable selection so as not to use unnecessary (overspecified) covariates.

For the SCA approach using PSM, we have argued that overspecifying the propensity score model should be of less concern than underspecification, since we only get the desirable properties from PSM when we control for all confounders. An underspecified model will in general lead to bias since we would then match with patients who are not representative to the treatment population with respect to confounding covariates. The procedure presented in section 4.1.2 suggests a method for correcting for such bias. However, we have not analytically derived any valid argument that this should be the case, and we therefore wish to investigate what happens when underspecifying the model in regard to confounding variables.

We considered the same simulated data sets as described in section 6.2, but having simulated 10 additional covariates with coefficients $a = b = c = 0$ in the data generating process. This means that specifying a model including any of these newly simulated variables results in an overspecified model. The $p = 20$ covariates were generated using the same distribution as described at the start of section 6.1.1, thus obtaining a common covariance structure between the covariates present in the data generating process and those that are not. Hence, including some of these variables could prove to be beneficial if the model is underspecified in regard to the 10 variables with corresponding nonzero values of a, b or c, since all covariates are to some extent correlated with the covariates in the data generating process.

In order to investigate what happens when overspecifying the models for the propensity and prognostic scores, we constructed these models based on all 20 covariates. To investigate the case of underspecification, the models were constructed based on 5 covariates included in the data generating process. In a similar way, we investigated the situation in which the models are both underspecified and overspecified in the sense of both including covariates that are part of and not part of the data generating process, which arguably mimics reality to the largest extent. This was carried out by training the prognostic and propensity score models based on 5 covariates in the data generating process as well as 5 covariates which were not.

In contrast to the results in section 6.2, raw covariate adjustment was done for only 2 covariates, which were included in the data generating process. We find that this situation mimics reality to a larger extent by only adjusting for a few number of covariates which are known to be highly prognostic, as described in section 2.3.2. In addition, we only considered the ANCOVA I estimators with design matrices (6.5) and (6.6), with the modification that m was trained on more covariates than the two that were directly adjusted for in the model.

On figure 6.3, we compare the performance using underspecified and overspecified prognostic and propensity score models, respectively, in the scenarios of a heterogeneous treatment effect with and without a covariate means shift in the historical data.

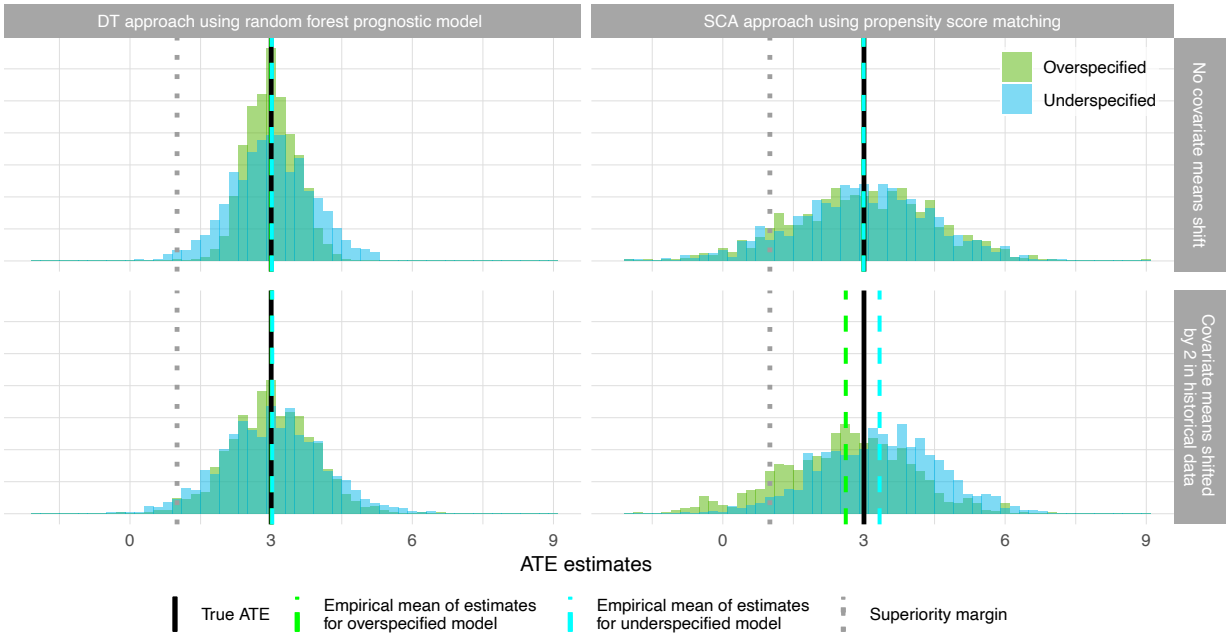


Figure 6.3: Empirical distributions of ANCOVA model ATE estimates obtained from $1,000$ simulated data sets under the heterogeneous treatment effect scenario and the scenario where the means of the distribution of the historical data covariates are shifted. Overspecified and underspecified models for the prognostic score and propensity score, using 5 additional covariates and failing to include 5 covariates in the data generating process, respectively.

For the approach based on digital twins, we see as expected that in both scenarios, underspecification has more severe consequences for the obtained efficiency than overspecification has. This result suggests that in practice, one should worry more about including too few than too many variables when training the prognostic model. However, no consequences in terms of bias occur from underspecification, since the approach is robust to biased prognostic models.

For the PSM approach, there is no clear indication regarding whether overspecification or underspecification is worst in terms of lost efficiency. This could be due to misspecification of the propensity score model in the first place, as previously discussed. The biased results for the scenario of the covariates being shifted in the historical data suggest that the method is unstable to

correcting differences between current and historical data when the model is either overspecified or underspecified. Overspecification should in principle only affect the variance of the estimated ATE and not the bias, as discussed in section 4.2. However, we also saw this tendency in the previous section when the model covariates were correctly specified, so the results could again be due to a misspecification of the propensity score model.

On figure 6.4, we investigate the realistic situation of simultaneously overspecifying and underspecifying the prognostic and propensity score models in the same scenarios as in figure 6.3.

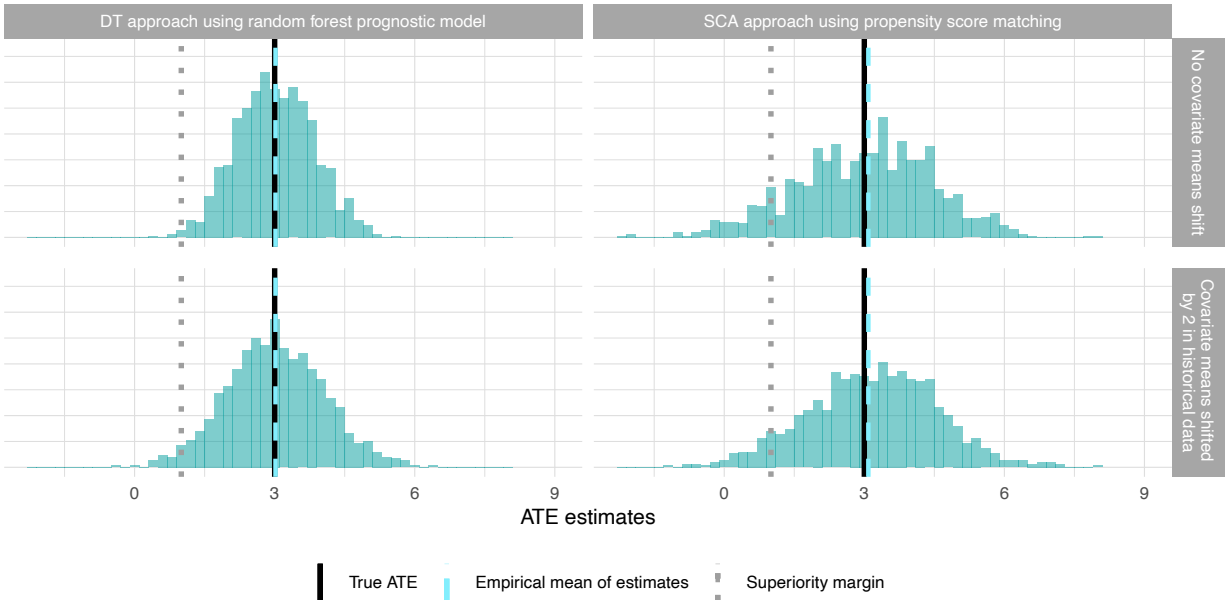


Figure 6.4: Empirical distributions of ANCOVA model ATE estimates obtained from $1,000$ simulated data sets under the heterogeneous treatment effect scenario and the scenario where the means of the distribution of the historical data covariates are shifted. Overspecified and underspecified models for the prognostic score and propensity score, using 5 additional covariates and failing to include 5 covariates in the data generating process.

For the DT approach, the results look much like the case of only underspecifying the model, as displayed in figure 6.3. This suggests that the random forest model is to some extent robust to overspecification, supporting the use of flexible, data-driven models when estimating the prognostic score in practice.

For the PSM method, the results are somewhat similar to the cases of only under or overspecifying the propensity score model. However, the bias induced from using only over or underspecified models seem to even out in this case, but in practice we cannot expect this to occur, making the method used here unreliable for use in practice. We conclude that the SCA approach using PSM remains prone to bias in estimating the ATE when misspecifying the propensity score model, whereas the DT approach provides an unbiased estimate of the ATE while reducing the variance, even when the prognostic model is misspecified in terms of model choice and/or covariate selection.

We investigated the RMSE of ATE estimates in all four scenarios in the realistic case of simultaneously overspecifying and underspecifying the propensity score and prognostic models. Results are displayed in figure 6.5. Interestingly for the purpose of using the methods in practice, the results were similar to those displayed in figure 6.2, apart from the problems of overfitting not being present since only a moderate number of covariates are adjusted for. Additional results are listed in table D.2, appendix D.2.

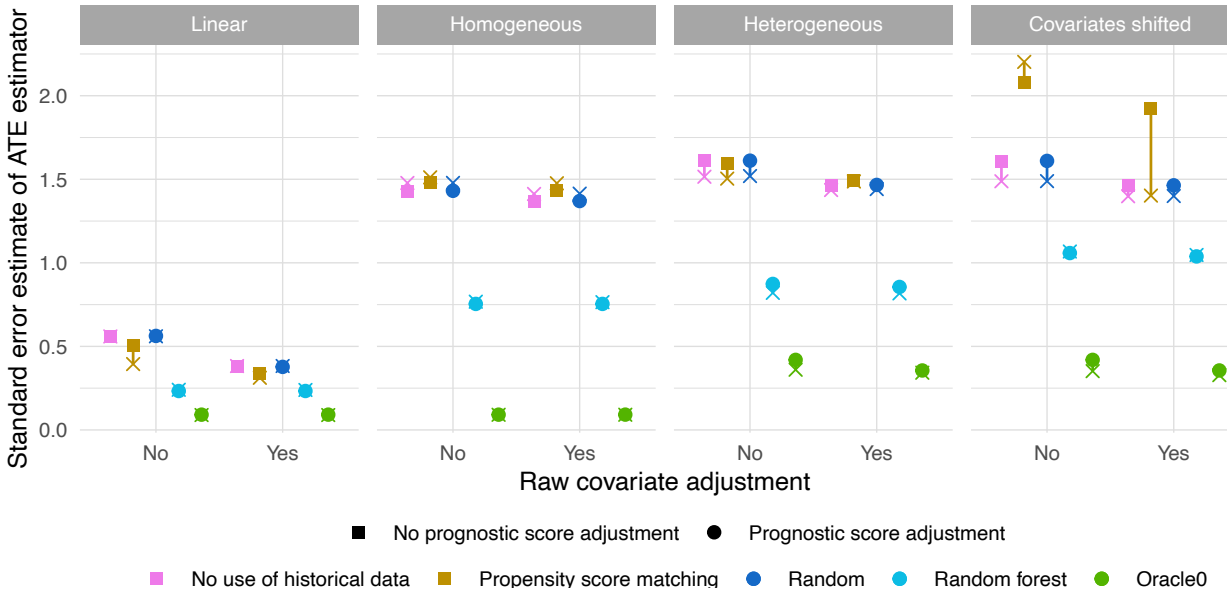


Figure 6.5: Standard error estimates (vertical axis) for AN(C)OVA model ATE estimates on $1,000$ simulated data sets under the four different scenarios (vertical panel classification). Filled points display empirical means of standard model dependent estimates across the $1,000$ simulated data sets. Crosses display the RMSE across the $1,000$ simulated data sets. Horizontal axis indicates whether raw covariate adjustments were included in the AN(C)OVA model for 2 of the 10 simulated covariates in the data generating process. No interaction terms between treatment allocation and raw covariates (as well as the estimated prognostic score for models "Random", "RF" and "Oracle0") were included in the AN(C)OVA model. Additional results are found in table D.2.

6.4 Varying Sample Sizes

We have seen from e.g. theorem 5.3.11 that if we provide a good prognostic model, the ANCOVA II estimator provides the asymptotically most efficient ATE estimator among all RAL estimators, at least when the treatment effect is homogeneous. In this section, we will investigate the finite sample properties of this estimator using different prognostic models together with PSM and standard AN(C)OVA estimators that do not leverage historical data. Specifically, we will investigate how fast we can increase power as $n, n' \to \infty$ and whether the individual methods provide control over the type I error. In practice, power needs to be estimated during the design phase of an RCT. Therefore, we will also evaluate methods for such prospective estimation.

We considered the heterogeneous treatment effect scenario according to table 6.1, where we used

the propensity score model and prognostic model which is both overspecified and underspecified, and where we utilise raw covariate adjustment for only 2 prognostic covariates, as described in section 6.3. We simulated $1,000$ data sets with $n = 2,000$, $n' = 10,000$, $n_0 = 800$ and $n_1 = 1200$. We then subsetted this large data set sequentially into 26 subsets, such that data sets with

$$n = 30, 40, 50, 60, 80, 100, 125, \ldots, 200, 250, \ldots, 500, 600, \ldots, 1000, 1200, \ldots, 2000 \quad (6.17)$$

and $n' = 5n$ were extracted, maintaining an allocation ratio of $r = n_1/n_2 = 1.5$ and such that participants in smaller data sets were contained in the larger data sets. This corresponds to increasing n and n' in a way that $n = \mathcal{O}(n')$ as specified in theorem 5.3.11.

In addition to the situation described above, where we increase n and n' simultaneously, we also consider situations where they are increased separately. Specifically, we extracted subsets in the way described above, with the exception that we keep a fixed number $n' = 5,000$ of historical data patients, in order to investigate the situation in which only the number of current trial participants increases with same rate described above. In the same way, we also extracted subsets such that $n_0 = 300$ and $n_1 = 200$ were fixed, while increasing the number of historical data patients n' with the same rate described above. These scenarios are both violations of the assumption in theorem 5.3.11 that both n and n' should increase such that $n = \mathcal{O}(n')$. Our purpose is to investigate the potential benefits from using digital twins, along with comparing this method with the SCA approach, when only increasing the number of historical data patients or current data patients. We relate the results of varying n and n' simultaneously to the results from varying them separately in the next section.

6.4.1 Performance Assessment

To evaluate the performance of the estimators as a function of increasing n and n' (simultaneously as well as separately), we will consider figures of the average ATE estimates and corresponding quantiles, the RMSE of the ATE estimate, the estimated power and type I error, and lastly the estimated L^2 norm of the prognostic models. We display the results of varying n and n' simultaneously in this section. Results relating to varying only n and only n', respectively, are displayed in appendix D.3.

In figure 6.6 the mean of the ATE estimates and their 25% and 75% quantiles are plotted as a function of n, where $n' = 5n$ for each method. Overall, the means of the ATE estimates seem to converge towards the true $\text{ATE} = 3$ for all the methods, but they vary in the rate of convergence. The two oracle estimators converge with the fastest rate. However, as already noted, the two oracle estimators are infeasible to determine in practice where the true data generating process is unknown. It is also seen that utilising digital twins with a random forest prognostic model converges faster than the ANOVA and ANCOVA methods.

For the random prediction model of the prognostic score, the mean ATE estimates align with the estimates derived from the ANCOVA method. This illustrates again that even when adding a

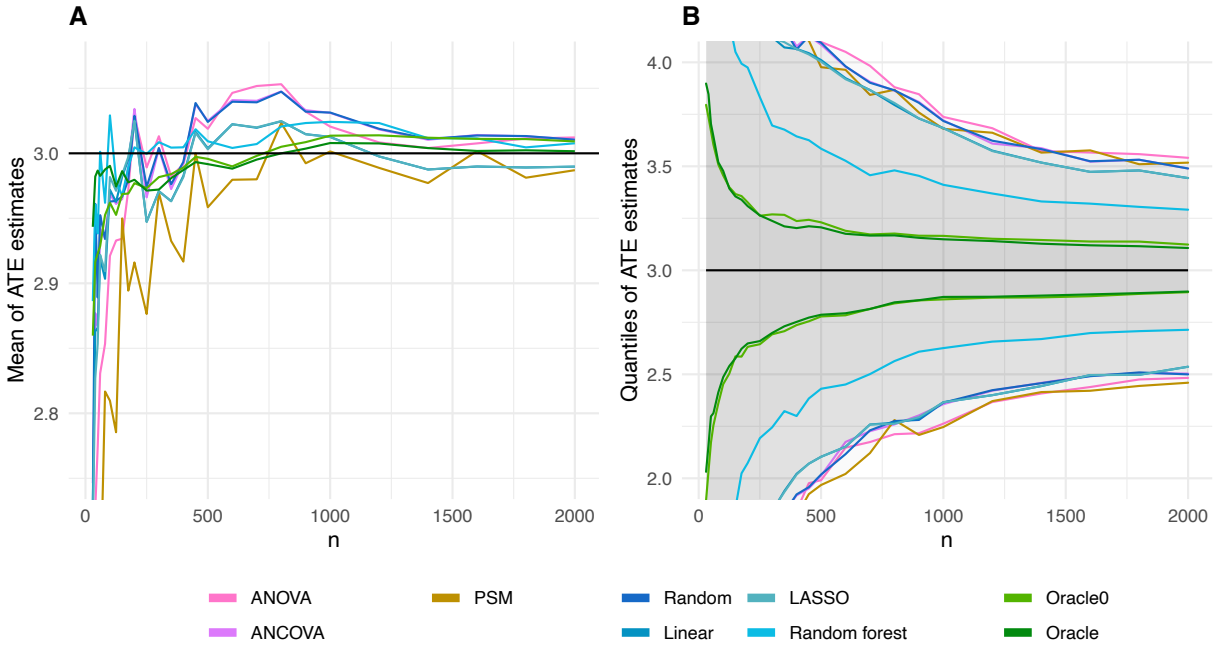


Figure 6.6: ATE estimates for all investigated models. Mean of ATE estimates (A) and their 25% and 75% quantiles (B).

prognostic score that does not correlate with the potential outcomes, and hence does not provide any efficiency gain, the efficiency of the ATE estimate is not worsened. This is supported by the quantile plot, where there is overlap between the quantiles of the ATE estimate for the ANCOVA method and the method using a random prognostic model. Looking at the quantile plot of the remaining DT estimators, we see a slight improvement over the ANCOVA model when using a linear or LASSO prognostic model, and we obtain a great improvement for the method using a random forest as a prognostic model. The improvements provided by the linear models are due to inclusion of covariates in the prognostic model which are not linearly adjusted for in the ANCOVA model. From this we again see that we have the largest gain when using a complex model that is able to capture the non-linearities in data, with the best models once again being the infeasible oracle estimators.

In figure 6.7 the RMSE of the ATE estimates are displayed as a function of n, where $n' = 5n$. We note that the y-axis is on a base 10 logarithmic scale. Similar to the results obtained in figure 6.6, the oracle estimators have the best performance, while the PSM only provides a slight decrease in RMSE compared to ANOVA, and still performs worse than the ANCOVA method. Again, there is not much to be gained from the random, linear or LASSO prognostic models, while the non-linear random forest model produces an efficiency gain. Since the difference between the RMSE curves are approximately constant on a logarithmic scale, the curves have an approximately constant scaling of each other by 10 to the power of the approximately constant difference.

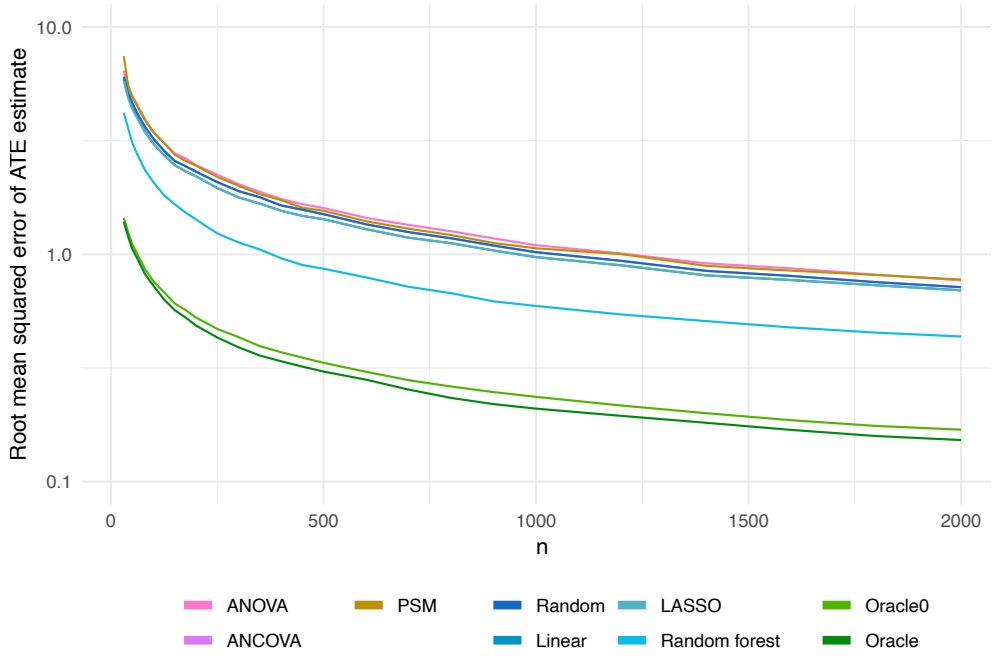


Figure 6.7: Root mean squared error of ATE estimates for each model in the heterogeneous scenario varying both n and n'.

In figure 6.8 the empirically estimated power is plotted in figure (A) and the empirically estimated type I error is plotted in figure (B) as a function of n, where $n' = 5n$. In regard to the empirically estimated power we can conclude similar results as for the previous plots, seeing that the two oracle estimators have the fastest convergence. The empirically estimated type I error is generally controlled using the DT methods, apart from the infeasible oracle estimator. For this estimator, the type I error is in general a little too large, again possibly being a result of overfitting. PSM gives an increase in power compared to the ANOVA method and the ANCOVA method for $n > 700$. However, the type I error is not controlled at the significance level of 2.5% for large n. The curves for the ANCOVA method and the method using a random prognostic score align, as expected from analytical results as well as from the previous analyses. The methods using a linear or LASSO prognostic model gives a slight increase in power compared to the ANOVA or ANCOVA method. Again, we have the largest feasible increase in power when we use the random forest model.

In figure 6.9 the empirically estimated L^2-norm of the difference between true and estimated prognostic scores is plotted as a function of n'. The linear and LASSO prognostic models do not seem to converge in L^2 towards the oracle estimator. This could explain why there is not a substantial increase in power or decrease in RMSE when we only vary n'. For the random forest prognostic model, there is a decrease in the L^2-norm, which again could explain why we see a large increase in power and large decrease in RMSE when we only vary n'. This illustrates the importance of a model that is complex enough to insure the convergence in L^2-norm.

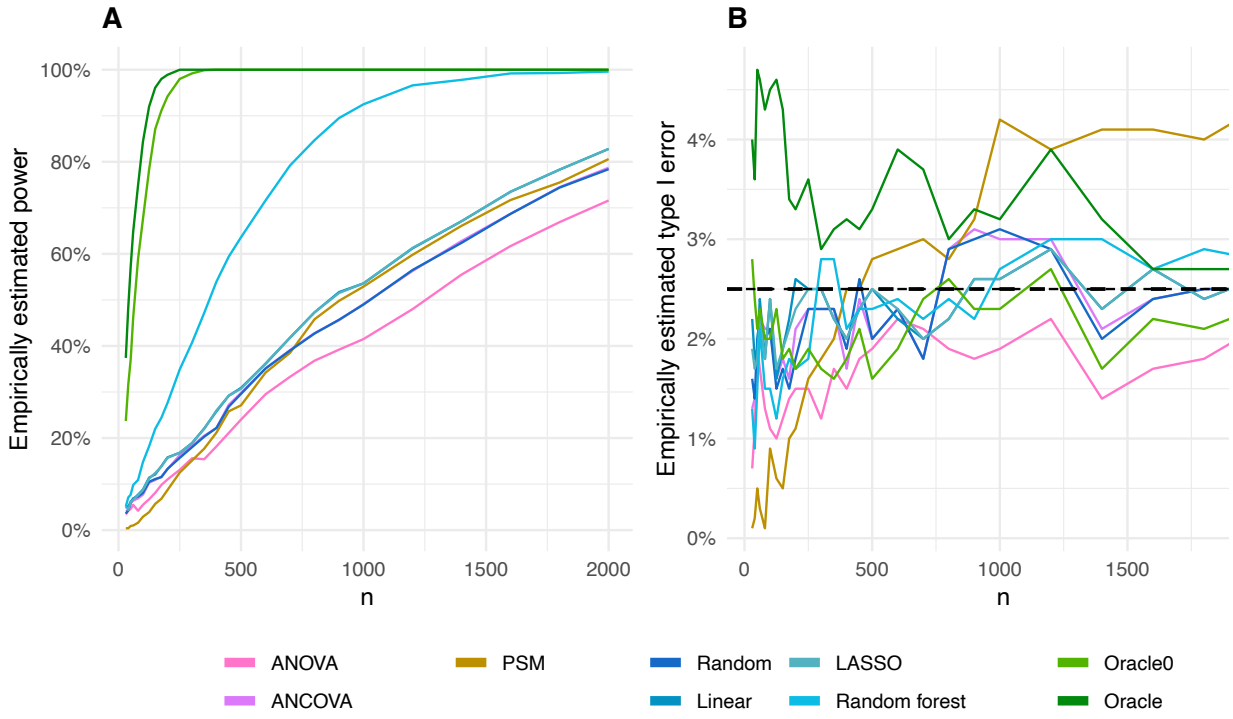


Figure 6.8: Empirically estimated power (A) and type I error (B) for each model in the heterogeneous scenario varying both n and n'.

Considering the results of varying n and n' separately, displayed in figure D.1 in appendix D.3, we see that the results of varying only n in figures A1-A4 look almost identical to the results in figures 6.6-6.8, indicating that n primarily controls the rate at which performance is increased as a function of increasing sample sizes, when fixing $n' = 5,000$. Only varying n' in figures B1-B4, we note that the results of methods not leveraging historical data are constant. Similarly, for the method using a random prognostic model, we have approximate constant results, since adding the random prediction to the ANCOVA model does not contribute with further efficiency gain. For the linear and LASSO models utilising historical data, we barely see a change in performance relative to the ANCOVA method.

The method utilising a random forest prognostic model seems to provide a substantial enhancement in performance as more historical data becomes available in regard to all of the performance measures. Specifically, on figure D.1(B3, B4) we only obtain a substantial increase in power when using the random forest prognostic model, while controlling the type I error rate. Therefore it seems that there is something to gain by using a digital twin model with sufficient complexity. Additionally, it seems that the gain to be exploited stagnates at around $n' = 5,000$, which is supported by figure 6.9, indicating that this size of the historical data should be sufficient. However, this might be different if more complex prognostic models are used or the data generating process is more complex than the one we investigate. As seen by figure D.1(B1) in

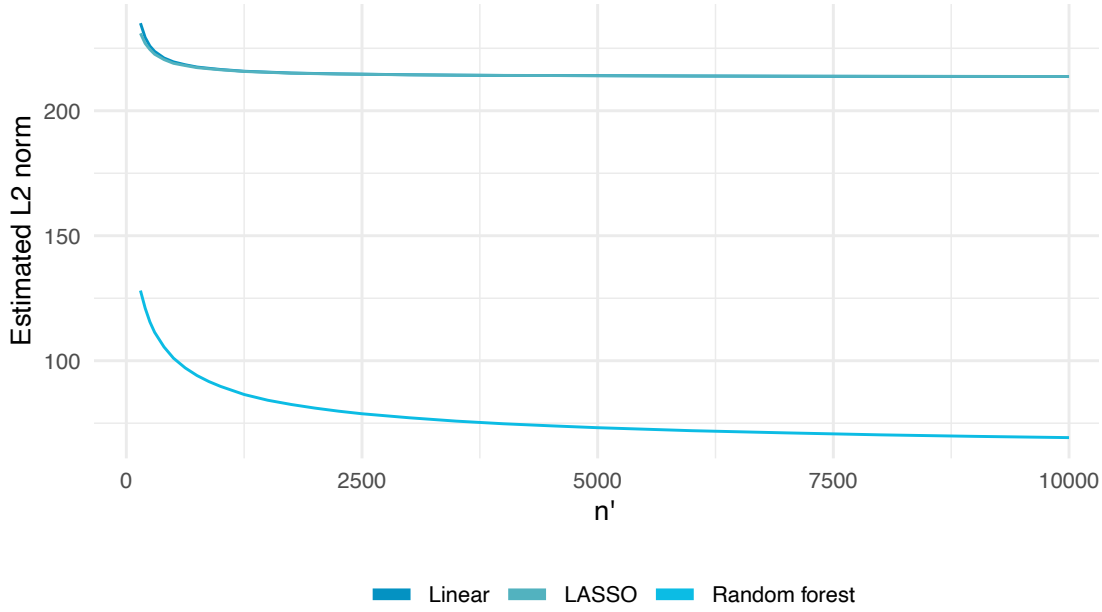


Figure 6.9: Empirically estimated L^2 norm of the difference between the estimated and true prognostic scores, for the digital twin models using a linear, LASSO and random forest prognostic model in the heterogeneous scenario, varying both n' for fixed $n = 500$.

appendix D.3, for the PSM method, the mean of the ATE estimate does not seem to converge towards the true ATE. This could indicate that our propensity score model does not utilise the added information that we should obtain for an increased n', supposedly due to misspecification of the propensity score model.

6.4.2 Prospective Estimation of Power from Sample Size

In the previous sections, we have evaluated the performance of different methods in terms of the empirically estimated gain in power. However, in practice a sample size calculation needs to be conducted prior to collecting data for the RCT. In chapter 3 and in section 5.4, we have provided methods for such prospective estimation of the required sample size for obtaining a desired power. In this section, we wish to evaluate whether these methods provide reasonable estimates for the required sample size when comparing the prospective estimates with the actual empirically estimated power.

We will start by considering the scenario of a heterogeneous treatment effect, using the values of current RCT sample size n investigated in the previous sections, meaning that we are in the same setup as described before and after equation (6.17). The derivations and (approximation) formulas in chapter 3 are based on the outcome variable having a certain distribution specified by the covariates. Hence, we are testing whether the methods for sample size calculation are robust to the realistic scenario of a partly misspecified model.

As pointed out in chapter 3, when conducting sample size calculations in practice, we would typically conduct sensitivity analyses in order to address the uncertainty inherent to estimating the necessary parameters. Since we will not delve into any methods for doing so here, we note that the results relate solely to evaluating the formulas for estimating the obtained power for some n and not how to address this uncertainty. However, in order to get a sense of the this uncertainty, we extracted data-driven 95% confidence intervals as the 2.5% and 97.5% empirical quantiles of the $1,000$ power estimates for each n. We note that this uncertainty relates only to estimation of the necessary parameters when we assume that the model is correctly specified.

We carried out sample size calculations for the ANOVA, ANCOVA, PSM and DT ANCOVA model utilising a random forest prognostic model based on the methods described in section 6.1.5. Based on these calculations, we found that the approximations based on the non-centrality parameter as well as the approximation formulas provided very similar results. In figure 6.10 we have displayed the empirically estimated power together with 2.5% and 97.5% quantiles of the prospective power estimates obtained from the Guenther-Schouten approximation, which is available for all the investigated models.

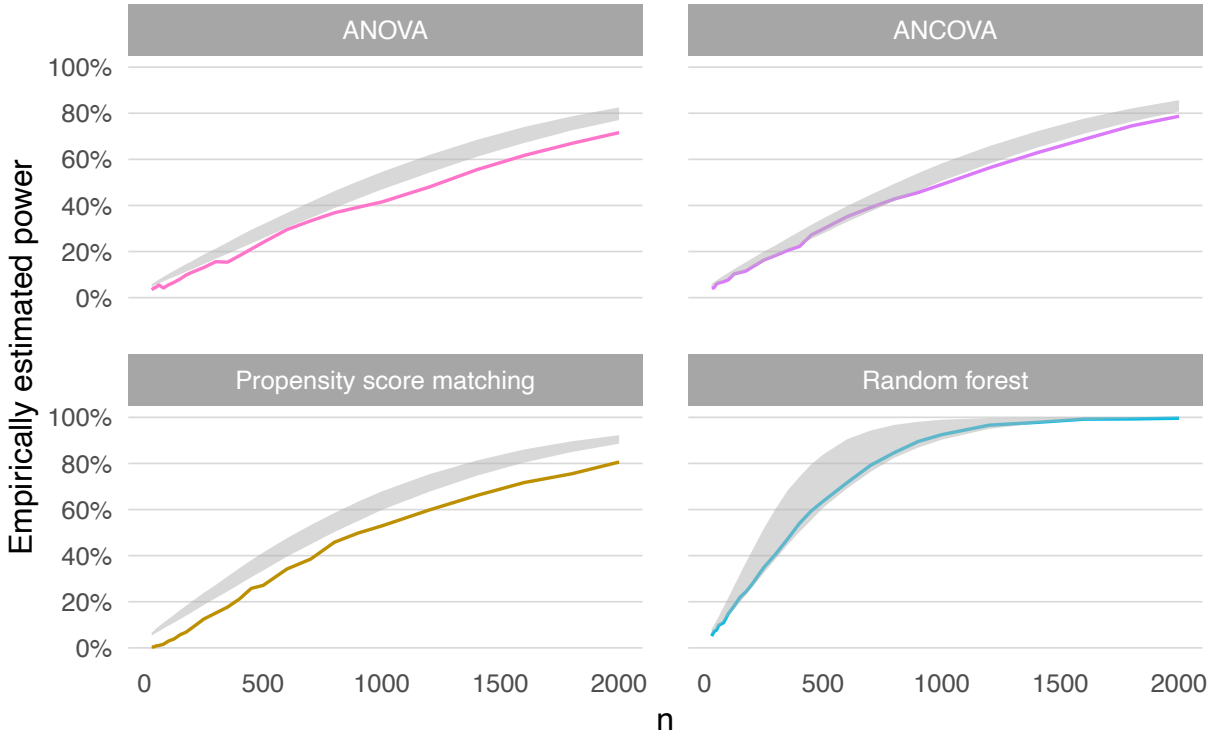


Figure 6.10: Empirically estimated power together with the mean and the 2.5% and 97.5% quantiles of the prospective Guenter-Schouten power approximation in the case of a heterogeneous treatment effect.

For all the methods the power is overestimated using the formulas derived in chapter 3. However, for the method using the random forest for estimating the prognostic score, the empirically estimated power lies just above the 2.5% quantile. The optimistic results of the prospectively

estimated power can be explained by the derivations in chapter 3 relying on the assumption that we have correctly specified the AN(C)OVA models. Specifically, the assumed model does not include interaction effects with the treatment assignment. However, our data is simulated using a heterogeneous treatment effect by having interaction effects in the data generating process. In practice, this means that the estimated correlation between the estimated prognostic score and the outcome is possibly overestimated for patients in the treatment group, resulting in optimistic estimation of power. Indeed, when we tested the Guenther-Schouten approximation in the homogeneous scenario, the power estimation seemed more reasonable for all four methods apart from the PSM method, as seen on figure D.2 in appendix D.4. Thus, a violation of the assumption of homogeneity seems to give too optimistic sample size approximations, hence accentuating the need to conduct sensitivity analyses when using these in practice.

7 Case Study – Digital Twins in Clinical Trials for Type 2 Diabetes

In this chapter we will examine the use of digital twins in clinical trials involving patients diagnosed with type 2 diabetes. Specifically, we wish to investigate the benefits of using digital twins on real world data in regard to decreasing the required sample size while maintaining a prespecified desired level of power.

The analyses are carried out from data sets provided by Novo Nordisk A/S originating from three previously conducted RCTs. The participants of the clinical trials were all diagnosed with type 2 diabetes, which is a chronic disease that affects around 8.5% of adults world wide. The pathogenesis of type 2 diabetes consists of insulin resistance and beta-cell impairment resulting in a decreased insulin secretion. Type 2 diabetes patients do not produce sufficient amounts of insulin, hence these patients need insulin injections to keep control of their blood sugar levels [64, 65].

There exist two types of insulin for injection; basal and bolus. Basal insulin is used to keep blood sugar levels stable between meals, whereas bolus insulin is administered in connection with a meal or high blood sugar levels. For this reason basal insulin is a long acting drug, that is administered subcutaneously (normally once or twice a day) and afterwards absorbed slowly into the bloodstream. This yields a steady plasma concentration of glucose between meals or at night. Bolus insulin is a fast acting insulin that is rapidly absorbed into the bloodstream [66]. Additionally, there exist different types of oral antidiabetic (OAD) medication with different ways of action. Commonly a drug called Metformin is used for type 2 diabetics. This works by increasing the efficacy of insulin and decreasing the excretion of sugar from the liver [67].

We begin the chapter by presenting the provided data sets and how we have curated them in order to be fit for our desired analyses. We then describe the methods used for these analyses. Lastly, we present the results of our analyses, which gives an indication of the possible gain from using digital twins in terms of lowering the required sample size for maintaining the desired level of power in the specific RCT. All analyses were carried out using `R` version 4.1.1, and all code can be found here.

7.1 Data Sets Provided by Novo Nordisk A/S

In this section we will describe relevant aspects of the three trials from which the data sets originate. All three trials involved patients diagnosed with type 2 diabetes at least 26 weeks prior to the screening visit, and no included patients were insulin naive; all patients have previously

received insulin exogenously. All trials were phase III, multicenter, 1:1 randomised, active controlled RCTs.

In order to ensure an allocation ratio of $r = 1$, the subjects in all trials were randomised in accordance with a forced balance randomisation scheme. As seen in example 2.2.1, this has the risk of inducing dependency of the treatment allocation variable W between observations, which means that the observations (X, W, Y) are not completely independent. With the assumption in chapter 5 of independency between observations, we thus cannot expect the analytical results of digital twins to hold completely. However, with somewhat large n, we expect that the dependency induced by a forced balance randomisation scheme is negligible.

In all trials, the Hemoglobin A1c (HbA1c) blood level percentage of the participants was measured. The HbA1c level measures the long-term average blood sugar level, and thus diabetes patients have a higher HbA1c level. The efficacy of insulin medication can therefore be assessed by the degree to which treated patients have a lowered level of HbA1c towards the normal range compared to the control group.

7.1.1 Trial NN1218-3853

Trial NN1218-3853 was a 26 weeks multinational, double-blinded, parallel group trial with the aim to compare the efficacy and safety of bolus insulin FIAsp (faster insulin aspart) versus bolus insulin aspart, both in combination with once daily basal insulin glargine (IGlar) and OAD in form of metformin. The trial design is summarised in figure 7.1.

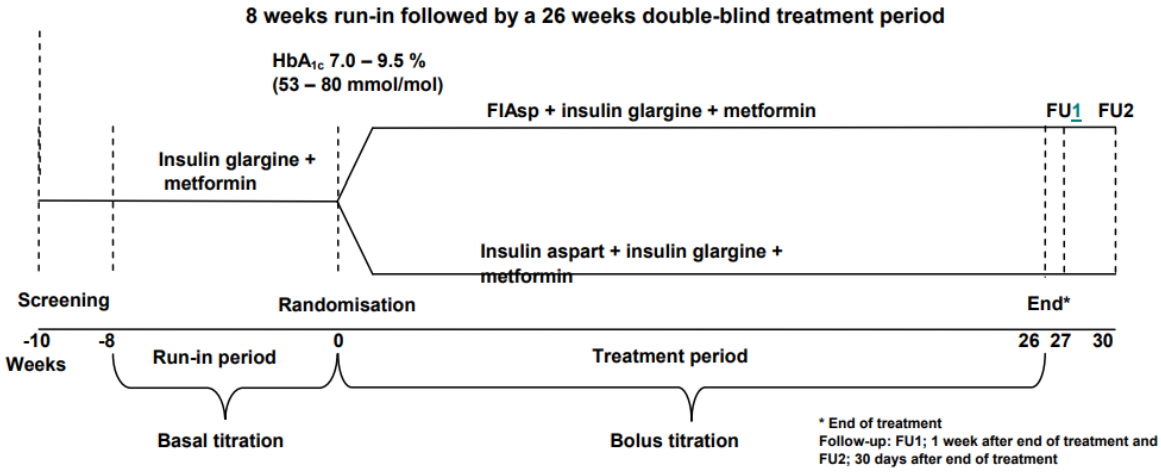


Figure 7.1: The trial design of trial NN1218-3853. Source: Clinical trial protocol (classified).

The total duration of the trial was approximately 40 weeks, including a 2 weeks screening period, 8 weeks run-in period, 26 weeks double-blinded treatment period and a follow-up period. At the screening visit subjects were assessed for their eligibility according to the inclusion and exclusion criteria listed in table E.1 found in appendix E.1. In the 8 week run-in period, eligible

subjects underwent basal insulin titration with insulin glargine and discontinued all OAD treatments except from metformin. All included subjects received metformin for at least 3 months prior to screening, and throughout the trial period the frequency and dosing with this drug should not be changed. At randomisation, subjects were randomised to receive either FIAsp (the treatment arm) or insulin aspart (the active control arm), both in addition to insulin glargine and metformin. At randomisation subjects fulfilled the randomisation criterion of having their HbA1c level between $7.0 - 9.5\%$, both inclusive.

According to planned treatment, the number of subjects in the control arm was 344, and the number of subjects in the treatment arm was 345. The primary objective of the trial was to confirm efficacy of FIAsp in terms of the glycaemic control assessed by change in HbA1c after 26 weeks from baseline. The trial was conducted as a non-inferiority trial with a non-inferiority margin of 0.4% and a significance level of 2.5%.

7.1.2 Trial NN1218-4049

Trial NN1218-4049 was a 18 weeks multinational, open-label, parallel group trial with the aim to compare the efficacy and safety of bolus insulin FIAsp in combination with basal insulin detemir, insulin glargine or human isophane insulin (NPH) versus only insulin detemir, insulin glargine or NPH, both in combination with metformin. The trial design is summarised in figure 7.2.

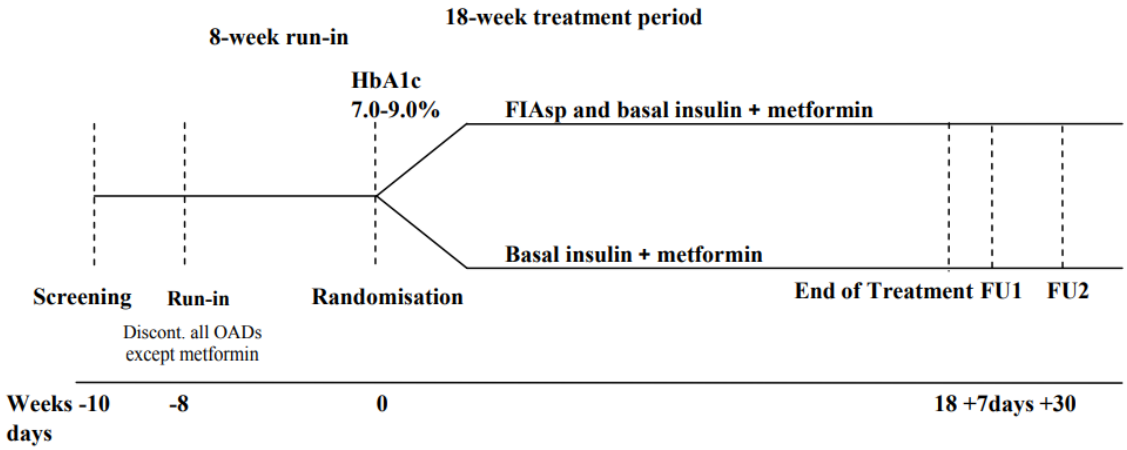


Figure 7.2: The trial design of trial NN1218-4049. Source: Clinical trial protocol (classified).

The total duration of the trial was approximately 32 weeks including a 2 weeks screening period, 8 weeks run-in period, 18 weeks treatment period and a follow-up period. At the screening visit subjects was assessed for their eligibility according to the inclusion and exclusion criteria listed in table E.2 found in appendix E.2. In the 8 week run-in period, eligible subjects underwent basal insulin titration with the insulin that the subject entered the trial with, and discontinued all OAD treatments, except from metformin. All subjects included received metformin for at least 3 months prior to screening, and throughout the trial period the frequency and dosing with this drug

should not be changed. At randomisation subjects were randomised to receive either bolus FIAsp (the treatment arm) or no bolus insulin (the control arm) in addition to insulin detemir, glargine or NPH in combination with metformin. At randomisation subjects fulfilled the randomisation criterion of having their HbA1c level between $7.0 - 9.0\%$, both inclusive.

The number of subjects with planned treatment being the control arm was 120 where the number of subjects with planned treatment being the treatment arm was 116. The primary objective of the trial was to confirm efficacy of FIAsp in terms of the glycaemic control assessed by change in HbA1c after 18 weeks. The trial was conducted as a superiority trial with a superiority margin of 0% and a significance level of 2.5%.

7.1.3 Trial NN1250-3998

Trial NN1250-3998 was a 64 weeks double blinded cross-over trial among patients from the United States of America, with the aim to compare safety of basal insulin degludec (IDeg) with basal insulin glargine (IGlar) both with or without OADs. The trial design is summarised in figure 7.3.

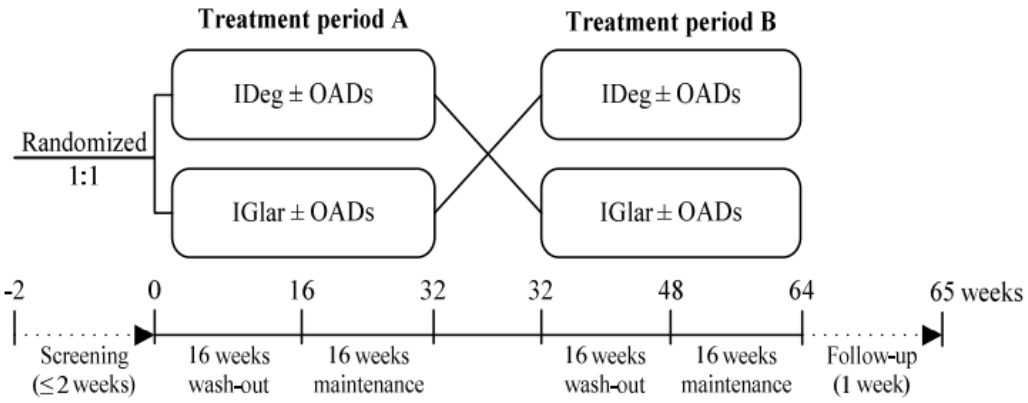


Figure 7.3: The trial design of trial NN1250-3998. Source: Clinical trial protocol (classified).

The total duration of the trial was approximately 64 weeks, starting with a 2 weeks screening period, which was followed by a randomisation visit, where each patient was randomised to a treatment sequence consisting of two treatment periods with either IDeg (the treatment arm) or IGlar (the control arm). Lastly, there was a 1 week follow-up period. In each treatment period, the patients started by going through a 16 weeks wash-out period. This period was added to avoid carry-over effects. In the wash-out period, the patients still received the treatment that they were allocated to, due to the fact that the objective of the trial was safety. At the screening visit, subjects were assessed for their eligibility according to the inclusion and exclusion criteria listed in table E.3 found in appendix E.3.

For both period A and B, the number of subjects with planned treatment being the control arm

was 361 where the number of subjects with planned treatment being the treatment arm was 363. The primary objective of the trial was to assess safety of insulin degludec in terms of showing a lower rate of severe or blood-glucose confirmed symptomatic hypoglycaemia compared to insulin glargine after 16 weeks of treatment. Even though efficacy was not the primary endpoint of this trial, the HbA1c was measured throughout the trial and therefore we can use the data set as historical data.

7.2 Curation of Data Sets

The data provided by Novo Nordisk A/S was delivered in a "raw" format in a standard called ADaM supported by the Clinical Data Interchange Standards Consortium (CDISC). Therefore we needed to determine which data sets contained the needed information, and then to curate these in order to use them for constructing digital twins. The overall process is described in section 7.2.1. In order to use the data sets for investigating the benefits of using the digital twin method, we also needed to split data between historical and current RCT data. This is described in section 7.2.2.

7.2.1 Standardising Trial-specific Data Sets

Even though data was provided in a specific standard, the same parameters were not all recorded across the three trials. Several data sets were associated with each trial. We chose to include data sets that contained the patients' concomitant medication, medical history, lab analyses, physical examinations, vital signs and a subject level data set that provides several baseline covariates in wide format. Most of the data sets were recorded in long format, so we converted the relevant covariates to wide format in order to extract one merged data set for each trial, which is standardised across the trials. For some of the data sets there was a large number of parameters, and including all of these could lead to overparameterisation causing a curse of dimensionality. For these data sets we constructed new parameters that summarise the information that the data set contained, as described in the following.

In the data set describing the physical examinations, different body regions categorised in nine different categories (the cardiovascular system, central and peripheral nervous system, gastrointestinal system incl. mouth, left eye ophthalmoscopy, right eye ophthalmoscopy, musculoskeletal system, respiratory system, skin, as well a category including head, ears, eyes, nose, throat and neck) were examined, and any clinically relevant abnormalities were recorded. Instead of including all nine parameters, we created a new parameter that counts the number of abnormalities in the physical examination. There were no missing values, so we did not induce any bias using this count variable.

For the two trials NN1218-3853 and N1218-4049, all patients were assessed for four prespecified comorbidities; diabetic nephropathy, diabetic neorpathy, diabetic retinopathy and macro angiopathy. These comorbidities were not severe enough to exclude the patients from the trials, but they are considered by medical professionals as clinically relevant. For this reason, we created a new

parameter that counts the number of these comorbidities in all patients. Specific comorbidities were also recorded in trial NN1250-3998, but the four clinically relevant comorbidities were not specified before the trial was conducted. For this reason we suspect that a possible surveillance bias could be present. Indeed, the total number of patients with at least one comorbidity was 32 (4.4%) in this trial, whereas the corresponding number was 385 (55.9%) in trial NN1218-3853, and in trial NN1218-4049 it was 87 (36.9%). However, this could also be due to other reasons, e.g. the trial in- and exclusion criteria differing between the trials.

We also considered including concomitant medication at baseline, but we chose not to include this, since the most important aspect is that patients are "in-control" when starting the study, which should be satisfied since all patients are not insulin naive. Furthermore, we control for baseline HbA1c which we expect to include any possible effect of which type of insulin the patients are on at treatment start.

Trial NN1250-3998 is the only trial that allowed for use of other OADs than metformin. In this trial, there was a total of 22 other OADs than metformin. We do not want to include too many parameters by using a factor indicating if the patient was on a specific OAD for all of the 22 other OADs. For this reason, we constructed one variable indicating if the patient was on metformin, and one variable indicating if the patient was on any other OAD.

For all of the trials, our outcome of interest will be the change from baseline until week 18 in HbA1c levels. For all trials, all randomised patients had a measurement of HbA1c at baseline. However, for some of the trials, the HbA1c level is not measured at week 18. For this reason, we examined when the HbA1c stabilises, so that previous or later measurements could be used for imputation. In figure 7.4, we have plotted the HbA1c measurements in each trial and each arm together with the mean HbA1c level within each arm. Based on the figure, we conclude that for all three trials, the HbA1c begins to stabilise around week 12 within both the control and treatment groups.

For trial NN1218-3853, we created the end of treatment HbA1c variable as the mean of the measurements at week 16 and 20. If one of these was missing, we used the latest measurement. If both were missing, we used the latest of the week 12 and 26 measurements. We chose the latest measurement as the first priority, since in accordance with figure 7.4, the HbA1c only starts to stabilise around week 12. All patients had a measurement at one or more of these weeks.

For trial NN1218-4049, we only had HbA1c measurements at week 12 and 18, so we used the latest measurement of these two weeks as the end of treatment measurement. Again, we chose the latest measurement, in accordance with figure 7.4. All patients had a measurement at one or more of these weeks.

For trial NN1250-3998 period A, we created the end of treatment HbA1c as the mean of the measurements at week 16 and 20. If one of these was missing, we used the latest measurement. If both were missing, we used the oldest of the week 24 and 28 or 32 measurements. All patients had a measurement at one or more of these weeks. We chose not to include data from period B, since we expect this data to be substantially different from the other trials, since the patients

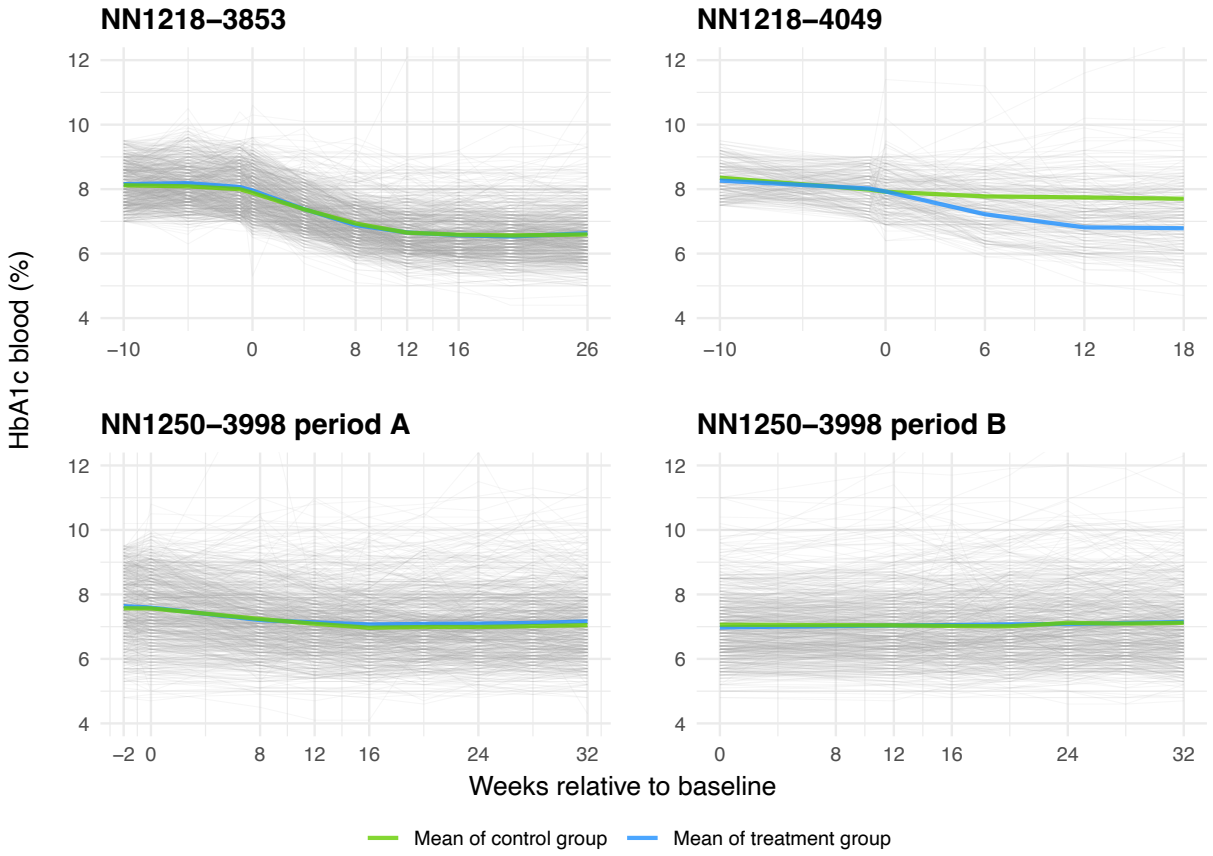


Figure 7.4: HbA1c measurements of all patients in the three trials, with the mean curve for the control arm (blue) and treatment arm (green).

in this period has already been part of a clinical trial for 32 weeks, which we expect to alter the results in a way that we cannot meaningfully account for.

Due to the fact that a lower level of the response corresponds to a positive effect, we will consider the change from baseline as $Y = Y^{\text{pre}} - Y^{\text{post}}$ to stay consistent with our assumption in the definition of the hypothesis tests in equations (3.9)-(3.11) that a positive effect implies a better outcome.

We then constructed a function that reads in the trial data and selects the baseline covariates we want to include. Then it filters the observations such that we use the baseline value or impute with the previously recorded values. This function also pivots the data from long to wide format, such that we have a column for each of the parameters we want to include. Within each trial, we used this function on each of the selected data sets and joined these to construct a standardised data set in wide format for each trial. For the cross-over trial, we did the same but only for period A. We only included patients that fulfilled the the inclusion, exclusion and randomisation criteria (recorded by a "full analysis set population flag" dummy variable in the provided data) and used the patient's planned treatment, thereby employing the intention-to-treat principle.

7.2.2 Current and Historical Data Sets

In order to examine the use of digital twins in a real data setting, we needed to split the data into historical data and current RCT data. We did so by defining our current RCT as a subset of the participants in trial NN1218-4049, while pooling the remaining patients from this trial with patients from trials NN1218-3853 and NN1250-3998 to form a pool of historical patients. Specifically, we used the subset of trial NN1218-4049 consisting only of patients receiving basal insulin glargine as our current RCT. Thus the current control arm is the part of trial NN1218-4049 receiving basal insuline glargine with metformin and without any bolus insulin, while the current treatment arm consists of the patients from NN1218-4049 that received bolus insulin FIAsp together with basal insuline glargine with metformin. We chose to do this in order to ensure that some historical data was to some degree representative of the current control arm, namely the subset of the glargine arms in trial NN1250-3998 that received metformin.

In appendix E.4 table E.4 an overview of all the covariates and the number of missing values are listed. We disregarded covariates that had a high proportion of missing values. After doing so, no missing values were present in the remaining covariates in the current trial. Missing values in the historical data were imputed by the `rfImpute` function, which for each observation iteratively imputes a proximity weighted average of the corresponding covariate values of the remaining observations, where the proximity is estimated based on a random forest procedure [68, 69]. The first imputation is done with mean imputation. Then, modeling each covariate as a function of the rest in a random forest, proximities are obtained for each pair of observations as the proportion of times that the observations are present in the same terminal node of the trees in the forest. The algorithm then iteratively imputes the covariate values of an observation as the weighted average of the corresponding covariate values of all other observations, using the proximities as weights. In general, uncertainty of imputation must be accounted for when estimating the variance of the ATE estimate. However, in our case, only the historical data contained missing values, affecting only the quality of the prognostic model, which we have already seen does not worsen the results compared to an ANCOVA model adjusting only for the raw covariates.

Thorlund et al. [70] discuss some important considerations in regard to the quality of the historical data in the context of the SCA approach to leveraging historical data. These considerations are important within this approach since the results obtained from external controls could be biased. As we have mentioned, this is not the case for the digital twins approach. However, we find some of the considerations relevant also in regard to constructing a prognostic model with good predictive performance.

Firstly, Thorlund et al. point out that the data collection process should be similar between the historical and current RCTs and that the investigated populations should be similar. Our three standardised data sets originate from RCTs, which means that the data collection in general adheres to a high level of stringency. This implies that the data collection process are highly similar, but we would expect higher similarity between the trials within the same project NN1218. We also examine similar populations since all three trials examines people with type 2 diabetes, and the clinical outcome and covariates are reasonably similar. The data quality is in general high for

RCTs compared to using observational studies or data from registries since the collection process in this case would be highly heterogeneous which could possibly lead to unknown confounding effects and missing data patterns that could bias the predictions made by the prognostic model. However, since our current RCT population originates from an open-label trial, the patients are not blinded to which treatment they receive, as opposed to the majority of the historical data, which consists of primarily double-blinded data. This could potentially induce a positive bias (in terms of the HbA1c level) for the prognostic model, when predicting outcomes of patients in the current RCT due to a possible placebo effect in the current data.

Secondly, Thorlund et al. state that the covariate distribution should be similar between the historical and current RCT data sets. There could be (and probably are) unobserved confounding covariates which are not reported in the two data sets. For this reason the eligibility criteria should be similar, which they are across the three trials, with only minor differences, as seen in tables E.1–E.3 in appendix E. However, this does not necessarily ensure that all important characteristics are similar across the patients. Furthermore, the outcome definitions should be similar between the historical and current RCT data sets, which they are for the two NN1218 trials. For trial NN1250-3998, the primary outcome was defined in order to assess safety and was thus not defined as the change from baseline HbA1c, but this was recorded for multiple weeks nonetheless, allowing us to regard this as our primary outcome of interest. Therefore we do not expect that this would alter the results.

In appendix E.4 we have displayed the empirical distributions of the continuous covariates (figure E.1) and the categorical covariates (figure E.2) in the current and historical data. For the continuous covariates, there are slight differences between the distributions of age, diabetes duration, baseline HbA1c, creatinine level, creatine clearence, sodium level and haematocrit level. However, these differences are very small, and we therefore expect that the continuous covariates in the historical data are representative for predicting the outcome on the control arm.

For the categorical covariates we see some larger differences. In the historical data, the nationality of subjects is primarily the United States of America, whereas the current RCT contains a large group of subjects from Argentina, India, Mexico and Romania, and these groups are not well represented in the historical data. This may also be the reason why there are some discrepancies in regard to race and if patients are Hispanic or Latino, where we see a large proportion of Asians and Hispanics or Latinos in the current RCT but only a small proportion in the historical data. Regarding the number of comorbidities, it seems that there are more patients with multiple comorbidities in the current RCT compared to the historical data. As noted earlier, this may be due to a surveillance bias for this covariate. Together, these discrepancies could indicate that the historical data is not completely representative of the control arm. Specifically, if there are inconsistencies in some strong predictors, this could skew the results of the prognostic model and also indicate that n' may not represent the number of historical data points that is representative of the current control group.

In table 7.1 we have summarised the number of patients in the current RCT and the historical data. Defining the historical control arm as the subset of historical data patients that received

the same treatment as the current control arm (basal insulin glargine and metformin), there were 237 patients that were directly representative of the current control arm. That is, only a minority of the historical data of size 1,492 was actually completely representative of the current control arm.

Trial	Number of patients
Current RCT	153
Current control arm	77
Current treatment arm	76
Historical data	1,492
Historical control arm	237

Table 7.1: Number of patients in the data split. The historical control arm consists of the patients in the historical data that received both insulin glargine and metformin and no bolus insulin.

It could be discussed whether or not we should use all of the historical data or just the subset that resembles the current control arm when training our prognostic model. One argument in favor of pooling the data is that we can expect the prognostic model to benefit from being trained on patients that are not completely representative of the current trial population, as long as we take the differences into account by including the corresponding covariates when training the model. Thus, we get more data to for our prognostic model to learn from, which is crucial in order to learn some of the complex non-linearities and interaction effects which are supposedly present in the underlying data generating process.

An argument against pooling is that when pooling, patients in the historical control arm do not resemble patients in the current control arm very well. As described earlier, this is a problem in SCA approaches like PSM, since the historical data would not necessarily resemble the current control and confounding covariates could thus skew the results. For the DT approach, the problem lies more in the fact that the model should learn from relevant data in order to possess a good predictive performance. The question is then whether pooling the data makes predictions based on the relevant subset that constitutes the current trial population better or worse, when the model is trained using a larger population.

Considering both the quality of historical data and the relative scarcity of data points if we were to only use highly similar patients from the historical data, we chose to use all historical data for training our model. An approach could have been to assess the predictive performance of the model trained among all historical data and the historical control arm, respectively, and the best performing model could be chosen. However, we found it convincing that using all historical data would provide a decent model, so we proceeded with this choice.

7.3 Methods

In order to estimate the ATE in the current RCT data, we performed analyses using both an ANOVA and ANCOVA model, that is, using the design matrices in equations (6.2) and (6.3), respectively, with $\mathbb{X}$ containing only the baseline HbA1c in accordance with regulatory guidelines. We then investigated how these models perform in comparison to a model utilising digital twins through a design matrix of the form given in equation (6.6) with the modification that only the baseline HbA1c was used as raw adjustment, and the prognostic model m was trained on all covariates (except trial variables) not marked with grey in table E.4 in the historical data.

7.3.1 Variance Estimation

ATE estimate variances were estimated using a heteroskedasticity-robust sandwich estimator, as described in section 3.5. In order to correct for the finite sample size, we used the HC_3 correction as suggested by Long and Ervin [34]. We did so even though there was no need for such robust variance estimation in our simulation study in the previous chapter, even in simulated scenarios of heteroskedasticity. This is due to working with real trial data in this case study, and thus, being unable to check the validity of different variance estimators, we wish to conduct our analyses in accordance with the opinions of regulatory authorities, which suggest to use such robust estimation [18].

7.3.2 The Prognostic Model

The prognostic model was specified as a random forest model trained using the default values $n'_{\min} = 5$ and $m = \lfloor p/3 \rfloor$ of the tuning parameters and $B = 500$, as described in section 6.1.3. Since random forests implicitly perform covariate selection through the splitting procedure, we trained the model using all covariates listed in table E.4 which are not marked with grey, except trial variables.

We suspected that since the baseline HbA1c value is a very strong predictor of the outcome, this covariate would too often be prohibited from being selected by the random forest procedure. We therefore investigated whether fine-tuning m could alleviate this problem. Specifically, we used 10-fold cross validation on the historical data to get an out-of-sample estimate of the prediction error of the prognostic models, measured by the MSE. However, fine-tuning m through the cross validation prediction error estimation, we did not obtain significant improvement in predictive performance. This could be explained by the fact that increasing m would lead to the trees in the forest becoming more similar, undermining the intention of the random forest to allow different trees to model unique data patterns without overfitting.

In order to accommodate the need to extensively use the baseline HbA1c while at the same time extracting useful information from the remaining covariates, we tried reducing the dimension of data through a procedure inspired by partial least squares (PLS). According to the partial least squares procedure [62, pp. 251–260] [63, pp. 79–81], a prespecified number of orthogonal dimensions of data are extracted as linear combinations of the covariates that contain a large

variance while at the same time being correlated to the outcome. In the classical procedure, the outcome is then regressed linearly on these new features. In our procedure, we instead trained the random forest model on the extracted features. In this way, we hoped to extract dimensions of data which generally contained a large proportion of the baseline HbA1c (since it is highly correlated to the outcome), so that the random forest procedure is forced to use this covariate extensively, while being able to use information from other prognostic covariates and still model the relationship non-linearly. We determined the number of new features extracted in this way through 10-fold cross validation and found that 10 dimensions were optimal in regard to reducing the MSE.

For both these random forest prognostic models, we calculated the MSE on the historical (training) data using out-of-bag (OOB) estimation as well as on the current control patients of the current RCT data.

7.3.3 Post hoc Power Estimation

We performed a post hoc sample size re-estimation and re-analysis of the current trial to illustrate the benefits of adjusting for the estimated prognostic score in terms of maintaining the same power with a reduced sample size. In order to do so, we estimated the power using the four different models in the current RCT by the appropriate Guenther-Schouten formulas, as presented in section 6.1.5. In this regard, we will refer to the Guether-Schouten approximation used in the DT approach suggested in section 5.4 (where the additional linear adjustment for baseline HbA1c

Parameter	**Value**
Significance level (α)	2.5%
Superiority margin (Δ_s)	0
Allocation ratio (r)	1
Assumed effect size (ATE)	0.6
Estimated standard deviation ($\widehat{\sigma}_Y$)	0.922
Inflation of standard deviation	1
Estimated correlation with raw baseline HbA1c level ($\widehat{\rho}$)	0.455
Estimated correlation with prognostic score ($\widehat{\rho}$)	0.466
Estimated squared multiple correlation coefficient with raw baseline HbA1c level and prognostic score ($\widehat{R}^2$)	0.308
Deflation of correlation coefficients	0.9

Table 7.2: Parameters used for re-estimation of power in the current RCT.

level is disregarded) as a *conservative* power estimation, whereas we will refer to the power estimation derived from (3.63) (where the additional linear adjustment for the baseline HbA1c level is included) as a *non-conservative* power estimation.

Re-estimation of the power was carried out with the parameters specified in table 7.2. The first four parameters are chosen in accordance with the parameters specified in the NN1218-4049 trial protocol since our RCT data consists of a subset of this trial's data. Estimation of the standard error as well as the correlation between the outcome and the prognostic score was carried out using all 153 patients in the current RCT data. The inflation and deflation parameters were used in order to conservatively estimate the power, by multiplying them with the estimated σ^2 and ρ^2 (R^2 for the non-conservative estimate), respectively. These specific choices of inflation and deflation parameters are based on a similar recalculation of power in another phase III trial [71] carried out by Unlearn.AI [72, pp. 16–19] [35].

7.4 Results

In this section we will assess the method of using prognostic models using a pre-specified analysis of the change in HbA1c after 18 weeks by comparing the models described in section 7.3. In table 7.3 we compare the resulting ATE and standard error of the models, as well as the MSE of the prognostic models.

			MSE	
Model	**ATE estimate**	**ATE SD**	**Historical data**	**Current data**
ANOVA	1.026	0.150		
ANCOVA	0.973	0.137		
Random forest	0.967	0.136	0.640	0.917
Random forest using PLS	0.906	0.129	0.626	0.650

Table 7.3: SD: Estimated standard deviation of the ATE estimate; PLS: Partial least squares; MSE: Mean squared error. Models *Random forest (using PLS features)* specify the ANCOVA model adjusting for baseline HbA1c and the estimated prognostic score. The MSE was estimated for the prognostic model using out-of-bag (OOB) estimation for the historical (training) data, and estimated from the predictions of only control patients in the current RCT data.

From the table we see that the largest decrease in standard deviation is obtained by using a random forest model combined with the use of partial least squares. Using this method we also obtain the lowest MSE of the prognostic model predictions in the current control data, implying that this model predicts the current control data most accurately. The small decrease in standard deviation for the standard random forest model could be explained by the larger MSE measured

on the current control data. This could indicate the use of PLS is appropriate in this case since HbA1c is a highly prognostic factor in itself.

In figure 7.5, the Guenther-Schouten approximated power is plotted as a function of the current sample size n, assuming a fixed allocation ratio of $r = 1$ and with relevant parameters estimated from all 153 current trial patients. The prognostic score of the DT model is constructed from the random forest using PLS features.

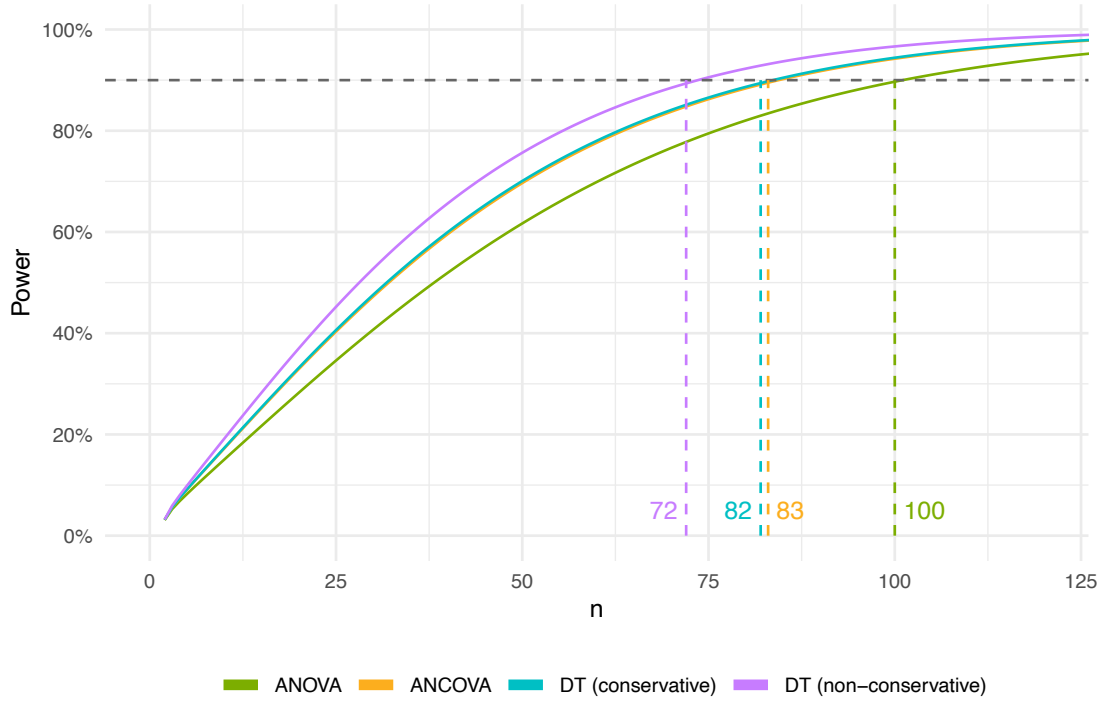


Figure 7.5: Post hoc Guenther-Schouten approximations of power obtained from the three different models for ATE estimation, based on the current RCT data. Horizontal dashed line is placed at 90% power, and the vertical lines indicate the sample size that gives an estimated power of 90%.

From the figure, we see that using the digital twin approach, it does seem possible to reduce the sample size while maintaining a desired level of power, especially when using the non-conservative estimate of the power. However, the gain from using the digital twin approach is in this case negligible if we use the conservative estimation of power.

We note that since this is a post hoc analysis, it serves only to illustrate the gain in power and is not directly applicable to how one could account for the prognostic scores in the design state of a trial. In practice, where the power has to be estimated prospectively, we can only use the historical data to estimate the entities needed in the prospective power estimation. The subset of historical control group patients eligible in regard to the current trial could be used for estimation of these parameters. However, since we had only a limited amount of historical control group data available, we found that the calculations would be too uncertain for such calculation to be

reliable. We note that in the case of prospective power estimation, the conservative Guenther-Schouten estimate may be more practically feasible than the non-conservative estimation, since more parameters need to be estimated in the latter approach, requiring more extensive considerations regarding sensitivity analysis. However, for the purpose of a post hoc analysis, one might argue that the non-conservative estimate of the power, albeit being more uncertain, is more appropriate, since the ANCOVA model adjusts for both the raw baseline value and the prognostic score, which should be taken into account.

According to the trial protocol, the desired power of the study was 90%, which is why figure 7.5 contains the estimated required number of patients for that power level, using the different models. According to the conservative estimate of the power using digital twins, the number of required patients in order to obtain the desired power can be reduced by 1 compared to raw adjustment for HbA1c, while the reduction is 18 compared to the ANOVA model where no adjustment is carried out. Using the non-conservative estimate of the power gain from (additional) adjustment for the prognostic score, the corresponding numbers of the reduction in required sample size are 11 and 28.

In order to test the validity of the approximated number of participants needed for a power of 90% under the three models, we randomly selected 72, 83 and 100 patients, maintaining an allocation ratio of $r = 1$, and re-estimated the ATE and the power among these patients using the ANOVA, ANCOVA and digital twin models, respectively. The non-conservative Guenther-Schouten approximation was used for estimating the power for the DT approach model. In order to minimise differences due to random variation, we selected patients such that patients contained in the smaller samples were included in larger samples. The results are summarised in table 7.4

Model	**Estimated power**	**Sample size**	**ATE estimate**	**ATE SD**
ANOVA	0.897	100	0.730	0.185
ANCOVA	0.909	83	0.611	0.172
Digital twin model	0.921	72	0.630	0.169

Table 7.4: Post hoc Guenther-Schouten approximation of power of the different models obtained from the estimated reduced sample size necessary for obtaining an estimated power of 90%, according to figure 7.5, along with the estimated ATE and estimated standard deviation (SD) of the ATE estimate.

From the table, we see that it does indeed seem possible to use the digital twin approach to reduce the sample size while maintaining a desired level of power. This is seen since even though the sample size is reduced from 83 in the ANCOVA model to 72 in the digital twin approach model, the power stays approximately at a level of 90%.

8 Future Perspectives on the Digital Twins Approach

In this chapter we will discuss several regulatory considerations in regard to using digital twins and reflect on possible further developments of the method as well as briefly discussing some alternative Bayesian approaches for using historical data, some of which have already been used for approval of drugs under specific circumstances. We will begin by discussing a recent draft qualification opinion on the use of digital twins in phase II and III RCTs published in March 2022 by EMA [72]. Here we will go through some of the specific concerns that were raised and relate some of these to relevant aspects of our simulation study conducted in chapter 6 and our case study conducted in chapter 7. We will then discuss some further developments on the digital twin method in the form of reflections on possible further improvements and expansions of the digital twin approach.

8.1 Regulatory Considerations

The 2022 EMA draft qualification opinion relates to the use of prognostic covariate adjustment (PROCOVA^{TM}) as a response to Unlearn.AIs application for qualification of the method from the Committee for Medicinal Products for Human Use (CMPH). In their application, Unlearn.AI focused on how the method provides a substantial decrease in standard deviation while controlling the type I error rate regardless of the specific prognostic model since the method is a special case of ANCOVA. They stated that the method is optimal when the prognostic model is highly correlated with the outcome, as we concluded from equation (5.85). Furthermore, they noted that there is a substantial dimensionality reduction by using a prognostic score in the ANCOVA model instead of raw adjustments, and that this makes prospective sample size calculation feasible since one only need to estimate the correlation between the prognostic model and the outcome and not a large number of population parameters, as seen in corollary 5.4.1. This is also consistent with the current EMA and FDA guidelines on covariate adjustment, as described in section 2.3.2. In this way it does not pose any additional statistical risk over any other pre-specified adjusted analyses. The dimensionality reduction also eliminates the possible problem of overfitting, which we observed in our simulation study in section 6.2, resulted in an increased type I error.

Regulatory standards provided by FDA and EMA often require drugs to have substantial evidence in regard to effectiveness from well-controlled trials [73, 74]. Therefore, it seems promising for the use of the digital twin approach that overall, the CMPH's response to the application was positive only with minor concerns, many of which were resolved by Unlearn.AI in a statistical handbook [75]. The CMPH agreed that the approach of adjusting by a prognostic score would provide an efficiency gain over other methods of adjustment. They generally agreed that under

the assumptions stated in theorem 5.3.11 the procedure can achieve a lower variance of the average treatment effect estimate, if the prognostic score is correlated with the outcome. Based partly on a simulation study in many ways similar to the one conducted in chapter 6, they acknowledged that the prognostic covariate adjustment produced lower MSE even when the assumptions were not strictly fulfilled, and that the results were supported by an empirical application to existing data, as we saw in chapter 7.

That the method is a special case of ANCOVA was agreed upon by the CMPH but with minor comments. Specifically, in the design stage of the trial, a prognostic model should be developed and validated on external data, as well as specifying prognostic covariates in the protocol to directly adjust for in the analysis. There is also substantial differences in the sample size calculations between the standard approach and the method using prognostic scores. As discussed in chapter 3, the sample size estimation using a standard approach is based on correlation between the prognostic covariates included in the model and the outcome. However, most trials are planned conservatively without taking the gain in adjustment of prognostic covariates into consideration. The method using prognostic scores uses the estimated correlation between the prognostic model and the outcome. This means that when conducting a sensitivity analysis, the methods differ in the sense that uncertainty in the estimated correlation coefficient between the prognostic model predictions and the outcome should be accounted for through a deflation parameter. The CMPH agreed that this correlation coefficient should be estimated on a separate data set to avoid overestimation. However, this would also be necessary if raw covariate adjustments are accounted for when using a traditional ANCOVA model. In the analysis stage, the digital twin approach would further differ in regard to adjustment of the prognostic score. However, the CMPH agreed that the properties regarding bias and control of type I error rate are the same as for the ANCOVA method. In addition, they emphasized the large importance of the prognostic models being trained on data independent of the current trial data for these results to hold, which is a prerequisite for the method to work, as we saw in theorem 5.3.11.

The CMPH had some concerns in regard to the method. One concern was that, when conducting a stratified randomisation with the intent to perform a subgroup analysis, the method would introduce multicollinearity between the prognostic score and any raw covariate included in the model, including interaction terms if such terms are specified by the model. This multicollinearity could alter the coefficients of these terms. However, Unlearn.AI provided an answer to this concern stating that the average treatment effect estimate would still be unbiased in this case, but that estimates related to the subgroup effect(s) would indeed be biased. That is, the method using prognostic scores is not intended for such kinds of subgroup analyses; hence, another linear model without adjustment for the prognostic score should be used for estimating subgroup (interaction) effects. Furthermore, in the context of stratified randomisation it is the standard to stratify by strong prognostic factors in the randomisation. However, since the prognostic scores are based on a large set of variables, it is not practical to implement this in the randomisation. Unlearn.AI clarified that the prognostic score should not be used for stratification and that the trial statisticians should only consider the strongest prognostic covariates for stratified randomisation, while keeping in mind that the estimated parameter related to this covariate is biased.

Another concern raised by the CMPH was that the prognostic score is trained under the conditions in the control arm, which could possibly imply that the correlation between the prognostic score and the outcome is larger for patients in the control arm than the treatment arm. Thereby we could encounter unequal residual variance between the two groups possibly leading to an inflation of the type I error rate. Unlearn.AI suggested to alleviate this problem by using heteroskedasticity robust estimation of the ATE estimate variance, such as the sandwich estimator that we employed in the case study presented in chapter 7. The CMPH agreed to this. However, we also carried out the simulations in chapter 6 using robust estimation, but we found that this did not change the results by much. We suspect that robust estimation might have led to more reliable estimation of the standard deviation if the treatment effect had been more heterogeneous and the allocation ratio remained different from 1. Furthermore, in regard to the prognostic model, the CMPH noted that outliers in the data used to fit the prognostic model may be influential points. Therefore it was recommended by Unlearn.AI that the prognostic model should be supported by model diagnostics to assess the robustness of the model in regard to large deviations of single observations.

It could be expected that the current RCT data could be incomplete in regard to some of the covariates included in the prognostic model. To handle such a problem, a missing data scheme which only depends on baseline covariates should be prespecified. Also the correlation between the outcome and the prognostic model may be expected to decrease when some important covariates are frequently missing, and this should be taken into considerations when conducting a sensitivity analysis.

Additionally, the CMPH raised the concern that the method should provide a substantial advantage over ANCOVA with a few covariate adjustments. As discussed in chapter 5 the gain is due to the non-linear effects captured by the prognostic model. That is, if the true data generating process does not contain many large non-linearities and interaction effects, the method would not provide large advantages. Unlearn.AI suggested that in the designing state of a trial one should determine the correlation between the outcome and a few highly prognostic factors in order to determine if the efficiency gain from using a prognostic model is considerable compared to using a standard ANCOVA or ANOVA. Furthermore, a guide on how to conduct a sensitivity analysis was also published by Unlearn.AI in the previously mentioned handbook [75].

Lastly, the CMPH noted that the sample size should be sufficiently large in order to have enough data on safety parameters or important subgroups, limiting the possible advantage of using a digital twin approach. This is especially a concern for large phase III studies where there are often requirements on the minimum number of subjects needed to be exposed to the drug. However, it is sometimes not feasible or ethical to use internal controls, for example in the case of paediatric or rare diseases. In these cases it can be necessary to include historical data. At the moment, no trial has been approved using the digital twin approach, possibly due to the novelty of the method. There are, however, examples of drugs that have been approved using historical data. One such example is described in the next section.

8.2 Further Developments

Research within the field of leveraging historical data is subject to rapid development. The use of digital twins in clinical trials is a relatively novel statistical method, and current research is conducted with the aim of improving the method. The most straightforward way to try and improve the method is by enhancing the performance of the prognostic model. This could be carried out by using more complex models in situations where sufficient data is available, which is the content of the following section. Secondly, in this thesis, we have restricted our attention to a frequentist approach for using digital twins. Other methods for leveraging historical data rely on Bayesian borrowing from previously conducted trials. The digital twins method can also be extended using a Bayesian approach, possibly gaining more power than the frequentist approach. Lastly, we shortly describe some loose considerations on possible future research on the digital twins approach.

8.2.1 Prognostic Models

The digital twin approach requires specification of a prognostic model. No matter the choice and predictive performance of this model, the approach ensures control of the type I error rate as long as the model is not excessively overspecified, as discussed in section 6.2. However, as equation (5.85) and our simulation study suggest, the performance of the method relies on the degree to which the prognostic model is able to exploit non-linear relationships in data to get strong predictive performance, with the oracle0 estimator providing an upper limit of the possible gain of the method. In our analyses on simulated and real world data, we restricted ourselves to consider the random forest model (in combination with the dimensionality reduction method of PLS), which has a non-comprehensive training procedure, having only a limited number of tuning parameters. However, much future work in improving the digital twin approach on real world data could revolve around constructing strong predictive models, which are sometimes often more complex in terms of fine-tuning parameters.

An example of another class of non-linear prediction model suitable for use as prognostic models include e.g. feedforward neural networks. Models within this class have the capability of modeling complicated correlations between the covariates and the outcome of interest in a data-driven way. This makes it a suitable prognostic model which should exploit the correlation between the covariates and the outcome that is beyond a purely linear relationship. The model can be fitted through the procedure known as backpropagation [76]. In general being an overspecified model which need estimation of a lot of parameters, risks of overfitting can be alleviated by regularizing the weight parameters directly or through drop-out learning, thus enabling modelling of complex relationships without overfitting [63, pp. 389–415] [62, pp. 403–458].

Another type of neural network well-suited for specification of the prognostic model is the class of models known as (conditional) restricted Boltzmann machines [77, 78]. This is a type of neural network, which is generative in the sense that it seeks to learn the joint probability distribution of all covariates. This learned distribution can then be used for sampling, thus generating a complete clinical record of the digital twin, including the outcome which we seek to predict. This model

would thus be suitable in case of wanting to perform multiple hypothesis tests, since we need a complex model that is able to predict several outcomes; particularly the covariates that we wish to perform tests for.

The (conditional) restricted Boltzmann machine has the possibility of being trained on historical data from repeated measurements at several time points for each patient, possibly exploiting a larger portion of information from each patient. Predictions can then be made from baseline values using a Markov chain structure where repeated measurements are drawn from learned conditional distributions [79]. This approach has been used to forecast the progression of Alzheimer's disease by Unlearn.AI [80]. The model can be trained by employing a maximum likelihood approach using the contrastive divergence procedure [81, 77]. Training can be improved by additionally including an adverserial model in the likelihood, so that the model is trained until this adversarial is unable to distinguish patients sampled from the model from patients sampled from the learned distribution [82]. Extending the (conditional) restricted Boltzmann machine with several hidden layers provide a more complex model possibly capable of modelling more complex relationships [83].

Increasing the sample size of historical data can enable the use of such complex prognostic models. A way of accelerating the performance of the prognostic models and thus the potential efficiency gain could be for pharmaceutical companies to share historical data for training prognostic models, which could lead to a large increase in sample size. However, in the context of borrowing data, considerations have to be made regarding data quality and homogeneity. Pooling data from multiple sources sets high requirements for professionals gathering and documenting data as well as professionals curating the data in accordance with the potential difference between data sources.

8.2.2 Bayesian Approaches to using Historical Data

In 2019, FDA approved a paediatric label expansion for intravenous Benlysta based on a study where historical data was used [10]. Benlysta was already an approved treatment for adults with active, autoantibody-positive systemic lupus erythematosus. Since paediatric lupus is very rare it was not feasible to conduct a fully powered trial and there are also several ethical problems in regard to exposing children to placebo. Furthermore, the disease shares the same pathophysiology and disease manifestations as in adults. For these reasons, the strategy was not to conduct any hypothesis testing but instead to show directional consistency with the results from the adult trial by using descriptive statistics.

The method relies on employing Bayesian borrowing of results obtained from adult patients from previously conducted phase III trials. Specifically, these trials are used as historical data to construct an "informative prior". Additionally, a (flat) "sceptical" prior, which does not attribute any belief to the historical data, is constructed. These two priors are then weighted in order to construct a mixture prior distribution, assigning some percentage of belief in the relevance of the treatment effect distribution obtained from adult patients. A likelihood constructed from the paediatric data is then assessed together with the mixture prior in order to construct a posterior

distribution of the treatment effect from which inference is carried out [10, 84].

FDA wanted the clinical team to examine how large the weight to the prior data should be before a "tipping point" where the results began to look convincing. In this case the tipping point was 55% weight on the prior data from the adult trials [10]. The FDA stated: "*Based on discussion and feedback obtained from the clinical team, it appears reasonable to assume at least 55% weight on the relevance of the adult information to the pediatric population and we can therefore conclude that there is at least 97.5% posterior probability that Benlysta has a positive treatment effect in pediatric subjects*" [85, p. 106].

This example shows how in some cases, it can be necessary to use novel statistical methods to get approval of a drug. Furthermore, it highlights some of the conditions that made approval possible, for example that the drug had already been shown to be effective in adults and that the disease had a similar pathophysiology in children as in adults. In the example, only a small control group was used due to ethical reasons and since the disease was very rare. For such rare diseases or for preliminary proof of concept studies, it is relevant to examine the use of methods for leveraging historical data for single arm trials. We have already peripherally touched upon the methods of propensity score matching as well as Bayesian borrowing as methods for doing so. With the digital twin approach, estimated prognostic scores could be used as actual outcomes of control group patients in single-arm studies to estimate the ATE directly from the average of individual patient differences between outcomes and prognostic scores, as described in equation 5.2. However, the problem with strictly controlling the type I error persists in any of these methods, originating from possible bias due to confounding effects not accounted for, since patients are not randomised.

In the context of the digital twin method, a less strict approach can be employed by taking a Bayesian point of view. Currently, Unlearn.AI are developing a Bayesian extension to the digital twin approach, namely the PROCOVA+^{TM} method. The method is an extension of the frequentist use of prognostic scores outlined in this thesis in the sense that the estimated prognostic scores are additionally treated as containing prior information on the true prognostic scores. This can be formalised as a prior containing information on parameters of the ANCOVA model used to estimate the average treatment effect. On one extreme, a strict belief can be taken in the estimated prognostic scores being true, thereby obtaining an ATE estimate based on the single-arm principle described in the previous paragraph, by adding the digital twins as synthetic control patients. On the other extreme, we can choose to employ no prior belief in the prognostic model, thereby obtaining the frequentist approach, retrieving strict type I error rate control. The Bayesian approach relies in interpolating between these two extremes in a way that power is gained and the type I error rate is limited to an acceptable degree. Walsh et al. [53] have shown that the Bayesian digital twin approach has allowed for further increase in power, but with a less strict control of the type I error rate. Furthermore, only a single "belief" parameter, which interpolates between the two extremes, needs to be specified for the method to work, and they suggest a method for choosing the prior based on the predictive performance on historical data. In addition, they have shown that under some reasonable conditions, for example that the belief parameter is chosen appropriately, the method limits the type I error rate [86, 72, 53].

8.2.3 Future Research

We have explored the digital twin approach restricted to RCTs with a continuous outcome, being relevant for efficacy trials. However, outcomes measured on other scales are typically also of interest. It seems plausible that similar advantages of the digital twin approach might be present for problems involving other types of response variables such as binary, time-to-event data or repeated measurements. This would entail that the method should be adjusted for the use in e.g. generalised linear models, survival models and mixed models for repeated measurements. Furthermore, the method could possibly be used for other estimands than ATE, possibly in other settings as well.

Another direction for future research could revolve around the case of a heterogeneous treatment effect; we have explored some results in this scenario, and we saw in the simulation study that the method seemed robust to heterogeneous effects, but further analytical results regarding this case remain to be investigated further.

9 | Conclusion

In this thesis, we have examined approaches for leveraging historical data in current randomised clinical trials (RCTs) with the aim to increase power while maintaining control over the type I error probability. We have compared a synthetic control arm (SCA) approach using the prevalent method of propensity score matching, commonly used in observational studies, with the novel statistical method of utilising digital twins to estimate the average treatment effect (ATE) in RCTs. Specifically, we considered using digital twins with different analysis of covariance (ANCOVA) model specifications to assess efficiency in the average treatment effect estimator measuring the efficacy of a medical intervention by a continuous outcome variable. Having laid out relevant theoretical aspects of randomised clinical trials, AN(C)OVA, sample/power calculations and the SCA approach using a propensity score matching (PSM) method adapted to the context of RCTs, we focused on the novel approach of digital twins within two-arm RCTs.

Using theory of influence functions, we were able to derive asymptotic distributions of different AN(C)OVA model ATE estimators within the digital twins approach. Specifically, we found that within RCTs, the practically infeasible oracle estimator provides the asymptotically most efficient ATE estimator among the large class of regular and asymptotically linear estimators. For each patient in the RCT, the oracle ATE estimator adjusts for the true conditional expected outcomes in the (hypothetical) scenario of both receiving and not receiving treatment, conditioning on the baseline covariates. When a homogeneous treatment effect is present, it suffices to employ the oracle0 estimator in order to get an asymptotically efficient estimator. The oracle0 ATE estimator adjusts only for the prognostic score, being the true conditional expected outcome had the patients been in the control group (which is the expected outcome of a digital twin).

In practice, estimates of the prognostic scores can be used in place of the true prognostic scores used by the infeasible oracle0 ATE estimator. Such estimates should be obtained from a prognostic model learned from independent historical data. We showed that under a homogeneous treatment effect, using a practically feasible ANCOVA ATE estimator, adjusting for an estimated prognostic score obtained from a reasonable prognostic model, provides an asymptotically efficient ATE estimate. That is, even though the true data generating process is not specified by a parametric ANCOVA model specification, this practically feasible model provides an asymptotically efficient ATE estimator within RCTs. Moreover, the gain obtained compared to raw covariate adjustment depends on the degree to which the prognostic model is able to exploit complex non-linear and interaction effects in the underlying data-generating process, calling for the use of machine learning models. In accordance with regulatory guidelines, linear relationships between strongly prognostic covariates are exploited to obtain larger power by raw covariate adjustment in the ANCOVA model.

We compared the use of the PSM method with different estimators utilising digital twins in a

simulation study. In estimating the ATE, we generally found that the method of PSM failed to increase power over the simple ANCOVA method, and in some cases the method, not being robust to misspecification of the propensity score model, lost control over the type I error rate. When the historical data had a different covariate distribution than the current RCT data, the PSM method introduced bias in the estimation of the ATE. Conversely, the digital twin method was able to maintain control over the type I error even when the prognostic model was biased and misspecified, and the distribution of the historical data was different than the current RCT data. Moreover, the method provided a gain in power compared to the ANCOVA method not leveraging historical data, even in cases when assumptions of analytical results were violated.

In the simulation study, we verified an important robustness property of the digital twins approach in regard to a worst case scenario of the performance of the prognostic model. Using a prognostic model estimating the prognostic score at random, we obtained practically the same results as a similar ANCOVA model not leveraging historical data. Using a machine learning random forest prognostic model provided a substantial gain in power over the ANCOVA model in multiple scenarios. In only one case of an excessively overspecified ANCOVA model, adjustment of estimated prognostic scores led to a slight increase in type I error rate, indicating a potential problem with using the digital twins method with overfitted models. However, regulatory guidelines prohibit overfitting of the ANCOVA model from occuring in practice, by allowing adjustments for only one or a few raw baseline covariates, and the problem of controlling the type I error rate is thus implausible to happen in practice. This was confirmed by all results from the simulation study mimicking a realistic scenario of both over- and underspecifying the prognostic model, and adjusting for only a few raw baseline covariates. Even in the case of a heterogeneous treatment effect combined with a covariate shift in the historical data distribution, the digital twin method performed better than ANCOVA models not leveraging historical data, thus gaining power while controlling the type I error rate.

By using real world data from three previously conducted RCTs provided by Novo Nordisk A/S, we found that the digital twins approach could successfully be employed to decrease the number of required participants in an RCT investigating the efficacy of bolus insulin FIAsp measured by change from baseline HbA1c. Specifically, we evaluated the performances of AN(C)OVA models with raw adjustment for baseline HbA1c as well as models additionally adjusting for the estimated prognostic scores using a random forest prognostic model, using features extracted with the partial least squares method. Based on a post hoc estimation of power, we found that keeping the power at the desired level of 90%, the digital twin approach with this random forest prognostic model managed to lower the required number of RCT participants to 72, compared to the 100 and 83 patients required in the ANOVA and ANCOVA models not leveraging historical data.

In practice, sample size calculations need to be carried out prospectively during the design phase of a trial, in order to obtain a desired level of power. In our simulation study, we verified the necessity of performing sensitivity analyses in order to account for potential violations of assumptions employed in our size calculation formulas. For the purpose of prospective power estimation using digital twins, Unlearn.AI have issued a handbook describing important aspects

associated with the procedure of using digital twins in RCTs. The handbook contained a comprehensive procedure for conducting conservative prospective sample size calculations within the framework of digital twins in the design phase of the current RCT, in order to be certain that the method does provide a power gain prior to conducting the RCT. The European Medicines Agency (EMA) recently issued a draft opinion on the method of digital twins in which usage of the method is encouraged when conducted in accordance with this handbook. The overall feedback from EMA was positive, focusing on the method potentially increasing power while controlling the I error probability.

Future improvements of the method of digital twins could be obtained from employing and fine-tuning various machine learning prognostic models capable of modelling non-linear and interaction effects, requiring a rich amount of relevant, curated historical data. Other current and future research areas include the use of Bayesian approaches with digital twins, as well as other estimands of clinical interest, requiring adaptation of the digital twin method to other models than the ANCOVA model.

10 | Bibliography

[1] *What Are Clinical Trials and Studies?* [Article]. National Institute on Aging; 2020 [cited 24/11/21]. Available at: https://www.nia.nih.gov/health/what-are-clinical-trials-and-studies.

[2] *Clinical Pharmacology 1: Phase 1 Studies and Early Drug Development* [pdf]. Gerlie Gieser, Ph.D. Office of Clinical Pharmacology, Div. IV; 2012 [cited 24/11/21]. Available at: https://www.fda.gov/media/84920/download.

[3] Schmidt B. *Proof of Principle studies*. Epilepsy Research. 2006;68(1):48–52.

[4] *Clinical Pharmacology 2: Clinical Pharmacology Considerations During Phase 2 and Phase 3 of Drug Development* [pdf]. KS Reynolds, Office of Clinical Pharmacology, Division 4; 2012 [cited 24/11/21]. Available at: https://www.fda.gov/media/84924/download.

[5] Imbens GW. *Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review*. The Review of Economics and Statistics. 2004;86(1):4–29.

[6] Sekhon JS. *The Neyman-Rubin Model of Causal Inference and Estimation Via Matching Methods*. In: Box-Steffensmeier JM, Brady HE, Collier D, editors. *The Oxford Handbook of Political Methodology*. Oxford University Press; 2008. p. 271–299.

[7] Austin PC. *An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies*. Multivariate Behavioral Research. 2011;46(3):399–424.

[8] Rosenberger WF, Lachin JM. *Randomization in Clinical Trials : Theory and Practice*. John Wiley & Sons; 2015. Available at: https://ebookcentral.proquest.com/lib/aalborguniv-ebooks/detail.action?docID=4042978.

[9] Schmidt C. *Pediatric Predicament*. Scientific American. 2017;317(3):24–25.

[10] *Bayesian borrowing of adult efficacy data in paediatric drug development: A Case Study* [pdf]. U.S. Food and Drug Administration; [cited 29/04/22]. Available at: https://www.fda.gov/media/152385/download.

[11] *Introduction to Causal Calculus* [pdf]. University of British Columbia: Sanna Tyrväinen; 2017 [cited 24/01/22]. Available at: https://www.cs.ubc.ca/labs/lci/mlrg/slides/doCalc.pdf.

[12] Pearl J. *Causality: Models, Reasoning and Inference (2nd ed.)*. Cambridge University Press; 2009. Available at: http://bayes.cs.ucla.edu/BOOK-2K/ch6-2.pdf.

[13] Kahan BC, Jairath V, Doré CJ, Morris TP. *The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies*. Trials. 2014;15:139.

[14] McCoy E. *Understanding the Intention-to-treat Principle in Randomized Controlled Trials.* The Western Journal of Emergency Medicine. 2017;18(6):1075–1078.

[15] Davison R, MacKinnon JG. *Econometric Theory and Methods*. New York, Oxford University Press; 2004. Available at: `http://qed.econ.queensu.ca/ETM/ETM-davidson-mackinnon-2021.pdf`.

[16] *Guideline on adjustment for baseline covariates in clinical trials* [pdf]. Committee for Medicinal Products for Human Use; 2015 [cited 30/11/21]. Available at: `https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf`.

[17] *ICH E9; Note for Guidance on Statistical Principles for Clinical Trials* [pdf]. Committee for Medicinal Products for Human Use; 1998 [cited 30/11/21]. Available at: `https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf`.

[18] *Adjusting for Covariates in Randomized Clinical Trials for Drugs And Biological Products - Guidance for Industry* [pdf]. U.S. Food and Drug Administration; 2021 [cited 30/11/21]. Available at: `https://www.fda.gov/media/148910/download`.

[19] Pocock SJ, Assmann SE, Enos LE, Kasten LE. *Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practive and problems*. Statistics in Medicine. 2002;21(19):2917–2930.

[20] Tsiatis AA, Davidian M, Zhang M, Lu X. *Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A pricipled yet flexible approach*. Statistics in Medicine. 2008;27(23):4658–4677.

[21] Kieser M. *Methods and Applications of Sample Size Calculation and Recalculation in Clinical Trials*. Springer; 2020.

[22] Wang B, Wang H, Tu XM, Feng C. *Comparisons of Superiority, Non-inferiority, and Equivalence Trials*. Shanghai Arch Psychiatry. 2017;29(6):385–388.

[23] Dumville JC, Hahn S, Miles JNV, Torgerson DJ. *The use of unequal randomisation ratios in clinical trials: A review*. Contemporary Clinical Trials. 2006;27(1):1–12.

[24] Wittes J. *Sample Size Calculations for Randomized Controlled Trials*. Epidemiologic Reviews. 2002;24(1):39–53.

[25] Kieser M, Wassmer G. *On the Use of the Upper Confidence Limit for the Variance from a Pilot Sample for Sample Size Determination*. Biometrical Journal. 1996;38(8):941–949.

[26] Guenther WC. *Sample Size Formulas for Normal Theory T Tests*. The American Statistician. 1981;35(4):243–244.

[27] Schouten HJA. *Sample size formula with a continuous outcome for unequal group sizes and unequal variances*. Statistics in Medicine. 1999;18(1):87–91.

[28] Tang Y. *Exact and Approximate Power and Sample Size Calculations for Analysis of Covariance in Randomized Clinical Trials With or Without Stratification*. Statistics in Biopharmaceutical Research. 2018;10(4):274–286.

[29] Frison L, Pocock SJ. *Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design*. Statistics in Medicine. 1992;11(13):1685–704.

[30] Zimmermann G, Kieser M, Bathke AC. *Sample Size Calculation and Blinded Recalculation for Analysis of Covariance Models with Multiple Random Covariates.* Journal of Biopharmaceutical Statistics. 2020;30(1):143–159.

[31] Schuler A. *Modern Causal Inference*. Notion; 2022. Available at: `https://zany-leech-f80.notion.site/Modern-Causal-Inference-32b4128f336d4e4794fce56d2de18ec5`.

[32] White H. *A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity*. Econometrica. 1980;48(4):817–838.

[33] MacKinnon JG, White H. *Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties*. Journal of Econometrics. 1985;29(3):305–325.

[34] Long JS, Ervin LH. *Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model*. The American Statistician. 2000;54(3):217–224.

[35] Schuler A, Walsh D, Hall D, Walsh J, Fisher C. *Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score*. arXiv. 2021;v3.

[36] Wang B, Ogburn EL, Rosenblum M. *Analysis of covariance in randomized trials: More precision and valid confidence intervals, without model assumptions*. Biometrics. 2019;75(4):1391–1400.

[37] *Multiple Endpoints in Clinical Trials: Guidance for Industry* [pdf]. U.S. Food and Drug Administration; 2017 [cited 25/01/22]. Available at: `https://www.fda.gov/files/drugs/published/Multiple-Endpoints-in-Clinical-Trials-Guidance-for-Industry.pdf`.

[38] *Guideline on multiplicity issues in clinical trials* [pdf]. Committee for Human Medicinal Products; 2017 [cited 25/01/22]. Available at: `https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-multiplicity-issues-clinical-trials_en.pdf`.

[39] Wiens BL. *A fixed sequence Bonferroni procedure for testing multiple endpoints*. Pharmaceutical Statistics. 2003;2:211–215.

[40] *Multiple Hypothesis Tests* [pdf]. James H. Steiger; 2015 [cited 01/02/22]. Available at: `http://www.statpower.net/Content/311/Lecture%20Notes/MultipleHypothesisTesting.pdf`.

[41] *Holm's procedure* [pdf]. Siva Balakrishnan; 2019 [cited 01/02/22]. Available at: `https://www.stat.cmu.edu/~siva/705/lec24.pdf`.

[42] Vickerstaff V, Omar R, Ambler G. *Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes.* BMC Medical Research Methodology. 2019;19(129).

[43] Lim J, Walley R, Yuan J, Liu J, Dabral A, Best N, et al. *Minimizing Patient Burden Through the Use of Historical Subject-Level Data in Innovative Confirmatory Clinical Trials: Review of Methods and Opportunities*. Therapeutic Innovation & Regulatory Science. 2018;52(5):546–559.

[44] Pocock SJ. *The combination of randomized and historical controls in clinical trials*. Journal of Chronic Diseases. 1976;29(3):175–188.

[45] Stuart EA, Rubin DB. *Matching With Multiple Control Groups With Adjustment for Group Differences*. Journal of Educational and Behavioral Statistics. 2008;33(3):279–306.

[46] Rosenbaum PR, Rubin DB. *The central role of the propensity score in observational studies for causal effects*. Biometrika. 1983;70(1):41–55.

[47] Gu XS, Rosenbaum PR. *Comparison of multivariate matching methods: Structures, distances, and algorithms*. Journal of Computational and Graphical Statistics. 1993;2(4):405–420.

[48] Austin PC. *Optimal caliper widths for propensity score matching when estimating differences in means and differences in proportions in observational studies*. Pharmaceutical Statistics. 2011;10(2):150–161.

[49] Abadie A, Spiess J. *Robust Post-Matching Inference*. Journal of the American Statistical Association. 2021;0(0):1–13.

[50] Cameron AC, Gelbach JB, Miller DL. *Robust Inference With Multiway Clustering*. Journal of Business & Economic Statistics. 2011;29(2):238–249.

[51] Austin PC. *Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score*. Pharmacoepidemiology and Drug Safety. 2008;17(12):1218–1225.

[52] Schuler A, Walsh D, Hall D, Walsh J, Fisher C. *Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score*. The International Journal of Biostatistics. 2021;(ahead of print).

[53] Walsh D, Schuler A, Hall D, Walsh J, Fisher C. *Bayesian prognostic covariate adjustment*. arXiv. 2020;v1.

[54] Hines O, Dukes O, Diaz-Ordaz K, Vansteelandt S. *Demystifying Statistical Learning Based on Efficient Influence Functions*. The American Statistician. 2022;(Ahead of print):1–13.

[55] *Semiparametric Statistics* [pdf]. Columbia University; 2018 [cited 18/03/2022]. Available at: `http://www.stat.columbia.edu/~bodhi/Talks/SPThNotes.pdf`.

[56] Tsiatis AA. *Semiparametric Theory and Missing Data*. Springer; 2006. Available at: `https://link.springer.com/content/pdf/10.1007%2F0-387-37345-4.pdf`.

[57] Hájek J. *A characterization of limiting distributions of regular estimates*. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete. 1970;14:323–330.

[58] *A note on semi-parametric estimators* [pdf]. Chen YC, University of Washington; 2020 [cited 18/03/2022]. Available at: `http://faculty.washington.edu/yenchic/short_note/note_EIF.pdf`.

[59] van der Vaart AW. *Asymptotic Statistics*. Cambridge University Press; 1998.

[60] Bao Y. *Should We Demean the Data?* Annals of Economics and Finance. 2015;16(1):163–171.

[61] Powell PD. *Calculating Determinants of Block Matrices*. arXiv. 2011;v1.

[62] James G, Witten D, Hastie T, Tibshirani R. *An introduction to Statistical Learning: With Applications in R*. Springer; 2021.

[63] Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer; 2017.

[64] *Hvad er type 2-diabetes?* [article]. Videncenter for Diabetes; 2021 [cited 23/09/21]. Available at: `https://videncenterfordiabetes.dk/viden-om-diabetes/type-2-diabetes/hvad-er-type-2-diabetes`.

[65] *Diabetes* [webpage]. WHO; 2021 [cited 20/05/22]. Available at: `https://www.who.int/news-room/fact-sheets/detail/diabetes`.

[66] *Basal- og bolusinsulin* [article]. Videncenter for Diabetes; 2022 [cited 20/03/22]. Available at: `https://videncenterfordiabetes.dk/viden-om-diabetes/type-1-diabetes/behandling/insulin/basal-og-bolusinsulin`.

[67] *Medicinsk behandling af type 2-diabetes* [article]. netdoktor.dk; 2020 [cited 20/03/22]. Available at: https://netdoktor.dk/medicin/typetodiabetes.htm.

[68] *Manual for Setting Up, Using, and Understanding Random Forest V4.0* [pdf]. Leo Breimann; 2003 [cited 22/04/22]. Available at: https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf.

[69] Breiman L. *Random Forests*. Machine Learning. 2001;45(1):5–32.

[70] Thorlund K, Don L, Park J, Mills E. *Synthetic and external controls in clinical trials – A primer for researchers*. Clinical Epidemiology. 2020;12:457–467.

[71] Quinn JF, Raman R, Thomas RG, Yurko-Mauro K, Nelson EB, Dyck CV, et al. *Docosahexaenoic acid supplementation and cognitive decline in Alzheimer disease: a randomized trial*. Journal of the American Medical Association. 2010;304(17):1903–11.

[72] *DRAFT Qualification opinion for Prognostic Covariate Adjustment* [pdf]. Committee for Medicinal Products for Human Use; 2022 [cited 28/04/22]. Available at: https://www.ema.europa.eu/en/documents/other/draft-qualification-opinion-prognostic-covariate-adjustment-procovatm_en.pdf.

[73] Jahanshahi M, Gregg K, Davis G, Ndu A, Miller V, Vockley J, et al. *The Use of External Controls in FDA Regulatory Decision Making*. Therapeutic Innovation & Regulatory Science. 2021;(55):1019–1035.

[74] *Choice of Control Group in Clinical Trials* [pdf]. Committee for Medicinal Products for Human Use; 2001 [cited 29/04/22]. Available at: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-10-choice-control-group-clinical-trials-step-5_en.pdf.

[75] *PROCOVA™ Handbook for the Target Trial Statistician* [pdf]. Unlearn.AI; 2021 [cited 28/04/22]. Available at: https://mcusercontent.com/bb49b1eeba8a691b82d19b68b/files/e80e25d0-52fa-0c58-dbd3-d031aa485ba9/UnlearnPROCOVAHandbookForTheTargetTrialStatistician.pdf.

[76] *Backpropagation* [article]. Brilliant.org; 2022 [cited 06/04/22]. Available at: https://brilliant.org/wiki/backpropagation/.

[77] Mnih V, Larochelle H, Hinton GE. *Conditional restricted Boltzmann machines for structured output prediction*. In: *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*. AUAI Press; 2011. p. 514–522.

[78] Fischer A, Igel C. *An Introduction to Restricted Boltzmann Machines*. In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer; 2012. p. 14–36.

[79] Walsh JR, Smith AM, Pouliot Y, Li-Bland D, Loukianov A, Fisher CK. *Generating Digital Twins with Multiple Sclerosis Using Probabilistic Neural Networks*. Cold Spring Harbor Laboratory; 2020.

[80] Fisher CK, Smith AM, Walsh JR. *Machine learning for comprehensive forecasting of Alzheimer's Disease progression*. Scientific Reports. 2019;9:13622.

[81] Hinton GE. *A Practical Guide to Training Restricted Boltzmann Machines*. In: *Neural Networks: Tricks of the Trade* (2nd ed.). Springer; 2012. p. 599–619.

[82] Fisher CK, Smith AM, Walsh JR. *Boltzmann Encoded Adversarial Machines*. arXiv. 2018;v1.

[83] Salakhutdinov R, Hinton G. *Deep Boltzmann Machines*. Proceedings of Machine Learning Research. 2009;5:1019–1035.

[84] *Bayesian dynamic borrowing for partial extrapolation and bridging studies: Methods & Case Studies* [pdf]. GSK; [cited 24/05/22]. Available at: `https://higherlogicdownload.s3.amazonaws.com/AMSTAT/87de5f9b-d337-4398-8c8e-a929aaace48b/UploadedImages/Mini_Conference/2021/01_Slides_Best.pdf`.

[85] *Multi-disciplinary Review and Evalaution Benlysta® (belimumab) for Intravenous Infusion in Children 5 to 17 Years of Age with SLE* [pdf]. U.S. Food and Drug Administration; [cited 29/04/22]. Available at: `https://www.fda.gov/media/127912/download`.

[86] *Applications of Digital Twins in Clinical Trials for Alzheimer's Disease* [pdf]. Unlearn.AI; 2021 [cited 29/04/22]. Available at: `https://www.unlearn.ai/post/applications-of-digital-twins-in-clinical-trials-for-alzheimers-disease`.

[87] Qiu Z, Guo W, Lynch G. *On Generalized Fixed Sequence Procedures for Controlling the FWER*. Statistics in Medicine. 2015;34(30):3968–3983.

A AN(C)OVA Model Derivations

This appendix contains derivations related to the ANOVA and ANCOVA models described in section 2.3.

A.1 ANOVA Model Maximum Likelihood Estimate

We wish to derive explicit expressions for the ML estimator of the ANOVA model in section 2.3. We will work with the ANOVA model parameterised as

$$Y_{wj} = (1-w)\beta_0 + w\beta_1 + \varepsilon_{wj}, \qquad j = 1, 2, \ldots, n_w, \quad w = 0, 1, \tag{A.1}$$

where $\varepsilon = [\varepsilon_{11}, \varepsilon_{12}, \ldots, \varepsilon_{0n_0}]^\top \sim \mathcal{N}\left(0, \mathrm{diag}_{n_1+n_0}(\sigma_Y^2, \sigma_Y^2, \ldots, \sigma_Y^2)\right)$ and the Y_{wj}'s are mutually independent. This reparametrisation from (2.17) corresponds to $\beta_W = \beta_1 - \beta_0$. Derivations could be carried out similarly for the model from the original parametrisation, but calculations are simplified by the reparametrisation. The the design matrix of the reparametrised model is given by

$$\mathbb{D} = \begin{bmatrix} 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{bmatrix}, \tag{A.2}$$

with the first column being indicator of being in the control group, and the second column being indicator of belonging to the treatment group. For a vector of the response variables $\mathbb{Y} = [Y_{11}, \ldots, Y_{1n_1}, Y_{01}, \ldots, Y_{0n_0}]^\top$, we now obtain the ML estimator as

$$\begin{aligned} \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix} &= (\mathbb{D}^\top \mathbb{D})^{-1} \mathbb{D}^\top \mathbb{Y} \\ &= \begin{bmatrix} n_0 & 0 \\ 0 & n_1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & \cdots & 0 & 1 & \cdots & 1 \\ 1 & \cdots & 1 & 0 & \cdots & 0 \end{bmatrix} [Y_{11}, \ldots, Y_{1n_1}, Y_{01}, \ldots, Y_{0n_0}]^\top \\ &= \begin{bmatrix} n_0^{-1} & 0 \\ 0 & n_1^{-1} \end{bmatrix} \begin{bmatrix} \sum_{w=1}^{n_0} Y_{0i} \\ \sum_{w=1}^{n_1} Y_{1i} \end{bmatrix} = \begin{bmatrix} \overline{Y}_0 \\ \overline{Y}_1 \end{bmatrix}. \end{aligned} \tag{A.3}$$

A.2 ANOVA Model Variance Estimate

Equation (A.3) implies that the usual unbiased variance of the MLE is given by

$$\widehat{\sigma}_Y^2 = \frac{||\mathbb{Y} - \mathbb{D}\widehat{\beta}||^2}{n_1 + n_0 - 2} = \frac{\sum_{w=0}^{1}\sum_{j=1}^{n_w(n)}\left(Y_{wj} - \overline{Y}_w\right)^2}{n_1 + n_0 - 2} = S_Y^2, \tag{A.4}$$

where the last equality follows directly from the expression of (3.25).

A.3 ANCOVA Model Maximum Likelihood Estimate

All derivations are based on [28]. We wish to derive explicit expressions of the MLE and the variance of the treatment estimate for the ANCOVA model presented in equation (3.38). First, we determine the design matrix as

$$\mathbb{D} = \begin{bmatrix} 0 & 1 & X_{11} \\ \vdots & \vdots & \vdots \\ 0 & 1 & X_{1n_1} \\ 1 & 0 & X_{01} \\ \vdots & \vdots & \vdots \\ 1 & 0 & X_{0n_0} \end{bmatrix}, \tag{A.5}$$

with the two first columns being indicators of belonging to the control- and treatment groups $w = 0$ and $w = 1$, respectively. This implies that

$$\mathbb{D}^\top\mathbb{D} = \begin{bmatrix} n_0 & 0 & n_0\overline{X}_0 \\ 0 & n_1 & n_1\overline{X}_1 \\ n_0\overline{X}_0 & n_1\overline{X}_1 & \sum_{w=0}^{1}\sum_{j=1}^{n_w} X_{wj}^2 \end{bmatrix}, \tag{A.6}$$

with $\overline{X}_w$ being the empirical mean of X for patients in group w. Since we have defined a normal linear model with response vector $\mathbb{Y}$ and design matrix $\mathbb{D}$, the MLE is given by $\widehat{\beta} = (\mathbb{D}^\top\mathbb{D})^{-1}\mathbb{D}^\top\mathbb{Y}$.

We now show that $(\mathbb{D}^\top\mathbb{D})^{-1} = AS_{xx}^{-1}A^\top + \text{diag}_3\left(n_0^{-1}, n_1^{-1}, 0\right)$ with $A = \begin{bmatrix}\overline{X}_0 & \overline{X}_1 & -1\end{bmatrix}^\top$ and $S_{xx} = \sum_{w=0}^{1}\sum_{j=1}^{n_w}(X_{wj} - \overline{X}_w)^2$. Since we have

$$AS_{xx}^{-1}A^\top = \frac{1}{\sum_{w=0}^{1}\sum_{j=1}^{n_w}(X_{wj} - \overline{X}_w)^2}\begin{bmatrix} \overline{X}_0^2 & \overline{X}_1\overline{X}_0 & -\overline{X}_0 \\ \overline{X}_1\overline{X}_0 & \overline{X}_1^2 & -\overline{X}_1 \\ -\overline{X}_0 & -\overline{X}_1 & 1 \end{bmatrix}, \tag{A.7}$$

we conclude that for $Q := -n_0\overline{X}_0^2 - n_1\overline{X}_1^2 + \sum_{w,j} X_{wj}^2$,

$$\begin{aligned}
&\mathbb{D}^\top\mathbb{D}\left(AS_{xx}^{-1}A^\top + \mathrm{diag}_3\left(n_0^{-1}, n_1^{-1}, 0\right)\right) \\
&= \frac{1}{\sum_{w=0}^{1}\sum_{j=1}^{n_w}(X_{wj} - \overline{X}_w)^2}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -\overline{X}_0 Q & -\overline{X}_1 Q & Q \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \overline{X}_0 & \overline{X}_1 & 0 \end{bmatrix} = I,
\end{aligned} \tag{A.8}$$

where the last equality follows since

$$\begin{aligned}
S_{xx} &= \sum_{w=0}^{1}\sum_{j=1}^{n_w}(X_{wj} - \overline{X}_w)^2 = \sum_{w=0}^{1}\sum_{j=1}^{n_w}\left(X_{wj}^2 + \overline{X}_w^2 - 2X_{wj}\overline{X}_w\right) \\
&= \sum_{w=0}^{1}\sum_{j=1}^{n_w} X_{wj}^2 + \sum_{j=1}^{n_1}\left(\overline{X}_1^2 - 2X_{1j}\overline{X}_1\right) + \sum_{j=1}^{n_0}\left(\overline{X}_0^2 - 2X_{0j}\overline{X}_0\right) \\
&= \sum_{w=0}^{1}\sum_{j=1}^{n_w} X_{wj}^2 + n_1\overline{X}_1^2 - 2n_1\overline{X}_1^2 + n_0\overline{X}_0^2 - 2n_0\overline{X}_0^2 = Q.
\end{aligned} \tag{A.9}$$

For a vector of the response variables $\mathbb{Y} = [Y_{11}, \ldots, Y_{1n_1}, Y_{01}, \ldots, Y_{0n_0}]^\top$, we now obtain the MLE as

$$\begin{aligned}
\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_X \end{bmatrix} &= (\mathbb{D}^\top\mathbb{D})^{-1}\mathbb{D}^\top\mathbb{Y} = \left(AS_{xx}^{-1}A^\top + \mathrm{diag}_3\left(n_0^{-1}, n_1^{-1}, 0\right)\right)\mathbb{D}^\top\mathbb{Y} \\
&= \left(S_{xx}^{-1}\begin{bmatrix} \overline{X}_0^2 & \overline{X}_1\overline{X}_0 & -\overline{X}_0 \\ \overline{X}_1\overline{X}_0 & \overline{X}_1^2 & -\overline{X}_1 \\ -\overline{X}_0 & -\overline{X}_1 & 1 \end{bmatrix}\begin{bmatrix} 0 & \ldots & 0 & 1 & \ldots & 1 \\ 1 & \ldots & 1 & 0 & \ldots & 0 \\ X_{11} & \ldots & X_{1n_1} & X_{01} & \ldots & X_{0n_0} \end{bmatrix}\right. \\
&\quad \left. + \mathrm{diag}_3\left(n_0^{-1}, n_1^{-1}, 0\right)\begin{bmatrix} 0 & \ldots & 0 & 1 & \ldots & 1 \\ 1 & \ldots & 1 & 0 & \ldots & 0 \\ X_{11} & \ldots & X_{1n_1} & X_{01} & \ldots & X_{0n_0} \end{bmatrix}\right)\mathbb{Y} \\
&= \left(S_{xx}^{-1}\begin{bmatrix} \overline{X}_1\overline{X}_0 - \overline{X}_0X_{11} & \ldots & \overline{X}_1\overline{X}_0 - \overline{X}_0X_{1n_1} & \overline{X}_0^2 - \overline{X}_0X_{01} & \ldots & \overline{X}_0^2 - \overline{X}_0X_{0n_0} \\ \overline{X}_1^2 - \overline{X}_1X_{11} & \ldots & \overline{X}_1^2 - \overline{X}_1X_{1n_1} & \overline{X}_1\overline{X}_0 - \overline{X}_1X_{01} & \ldots & \overline{X}_1\overline{X}_0 - \overline{X}_1X_{0n_0} \\ X_{11} - \overline{X}_1 & \ldots & X_{1n_1} - \overline{X}_1 & X_{01} - \overline{X}_0 & \ldots & X_{0n_0} - \overline{X}_0 \end{bmatrix}\right. \\
&\quad \left. + \begin{bmatrix} \begin{bmatrix} 0 & \ldots & 0 \\ n_1^{-1} & \ldots & n_1^{-1} \\ 0 & \ldots & 0 \end{bmatrix}_{3\times n_1} & \begin{bmatrix} n_0^{-1} & \ldots & n_0^{-1} \\ 0 & \ldots & 0 \\ 0 & \ldots & 0 \end{bmatrix}_{3\times n_0} \end{bmatrix}\right)\mathbb{Y}
\end{aligned} \tag{A.10}$$

$$= S_{xx}^{-1} \sum_{w=0}^{1} \sum_{j=1}^{n_w} \begin{bmatrix} \left(\overline{X}_w \overline{X}_0 - \overline{X}_0 X_{wj}\right) Y_{wj} \\ \left(\overline{X}_1 \overline{X}_w - \overline{X}_1 X_{wj}\right) Y_{wj} \\ \left(X_{wj} - \overline{X}_w\right) Y_{wj} \end{bmatrix} + \begin{bmatrix} n_0^{-1} \sum_{j=1}^{n_0} Y_{0j} \\ n_1^{-1} \sum_{j=1}^{n_1} Y_{1j} \\ 0 \end{bmatrix}.$$

Thereby, defining $S_{xy} \coloneqq \sum_{w=0}^{1} \sum_{j=1}^{n_w} \left(X_{wj} - \overline{X}_w\right) Y_{wj}$, we have $\widehat{\beta}_X = S_{xx}^{-1} S_{xy}$. Furthermore, we see that for $k = 0, 1$

$$\begin{aligned} \widehat{\beta}_k = S_{xx}^{-1} \overline{X}_k \sum_{w=0}^{1} \sum_{j=1}^{n_w} \left(\overline{X}_w - X_{wj}\right) Y_{wj} + n_k^{-1} \sum_{j=1}^{n_k} Y_{kj} &= -S_{xx}^{-1} \overline{X}_k S_{xy} + \overline{Y}_k \\ &= -\widehat{\beta}_X \overline{X}_k + \overline{Y}_k. \end{aligned} \tag{A.11}$$

Later, we will use that $\widehat{\beta}_X$ can also be expressed as

$$\begin{aligned} \widehat{\beta}_X &= S_{xx}^{-1} \sum_{w=0}^{1} \sum_{j=1}^{n_w} \left(X_{wj} - \overline{X}_w\right) \left(Y_{wj} - \overline{Y}_w\right) \\ &= \frac{\sum_{w=0}^{1} \sum_{j=1}^{n_w} \left(X_{wj} - \overline{X}_w\right) \left(Y_{wj} - \overline{Y}_w\right)}{\sum_{w=0}^{1} \sum_{j=1}^{n_w} (X_{wj} - \overline{X}_w)^2}, \end{aligned} \tag{A.12}$$

which is true since

$$\begin{aligned} \sum_{w=0}^{1} \sum_{j=1}^{n_w} \left(X_{wj} - \overline{X}_w\right) \overline{Y}_w &= \sum_{w=0}^{1} \overline{Y}_w \sum_{j=1}^{n_w} \left(X_{wj} - \overline{X}_w\right) \\ &= \sum_{w=0}^{1} \overline{Y}_w \left[n_w \overline{X}_w - n_w \overline{X}_w\right] = 0. \end{aligned} \tag{A.13}$$

Using equation (A.11), the estimated treatment effect is given by

$$\widehat{\beta}_1 - \widehat{\beta}_0 = \overline{Y}_1 - \overline{Y}_0 - \left(\overline{X}_1 - \overline{X}_0\right) \widehat{\beta}_X. \tag{A.14}$$

Tang [28] has shown that when X is p-dimensional, that is, $X = \left(X^1, X^2, \ldots, X^p\right)$, we get equivalently that

$$\widehat{\beta}_1 - \widehat{\beta}_0 = \overline{Y}_1 - \overline{Y}_0 - \sum_{l=1}^{p} \left(\overline{X}_1^l - \overline{X}_0^l\right) \widehat{\beta}_{X^l}. \tag{A.15}$$

A.4 (Estimated) Variance of the ANCOVA ATE Estimator

All derivations are based on [28] and [25]. We now wish to estimate the variance of this treatment effect estimator. First, we note that for the normal linear model,

$$\widehat{\beta} = \left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_X\right)^\top \sim \mathcal{N}_3\left(\beta, \sigma^2(\mathbb{D}^\top\mathbb{D})^{-1}\right), \tag{A.16}$$

giving that the estimated treatment effect is unbiased, and that

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(\widehat{\beta}\right) &= \sigma^2\left(AS_{xx}^{-1}A^\top + \mathrm{diag}_3\left(n_0^{-1}, n_1^{-1}, 0\right)\right) \\ &= \sigma^2\left(S_{xx}^{-1}\begin{bmatrix}\overline{X}_0^2 & \overline{X}_1\overline{X}_0 & -\overline{X}_0 \\ \overline{X}_1\overline{X}_0 & \overline{X}_1^2 & -\overline{X}_1 \\ -\overline{X}_0 & -\overline{X}_1 & 1\end{bmatrix} + \begin{bmatrix}n_0^{-1} & 0 & 0 \\ 0 & n_1^{-1} & 0 \\ 0 & 0 & 0\end{bmatrix}\right).\end{aligned} \tag{A.17}$$

From this covariance matrix, we can directly see expressions of variances and covariances between the control and treatment estimators given as

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(\widehat{\beta}_w\right) &= \sigma^2(S_{xx}^{-1}\overline{X}_w^2 + n_w^{-1}), \qquad w = 0, 1 \\ \mathbb{C}\mathrm{ov}\left(\widehat{\beta}_0, \widehat{\beta}_1\right) &= \sigma^2(S_{xx}^{-1}\overline{X}_1\overline{X}_0).\end{aligned} \tag{A.18}$$

Therefore we have

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(\widehat{\beta}_1 - \widehat{\beta}_0\right) &= \sigma^2\left(S_{xx}^{-1}(\overline{X}_1^2 + \overline{X}_0^2) + n_1^{-1} + n_0^{-1} - 2S_{xx}^{-1}\overline{X}_1\overline{X}_0\right) \\ &= \sigma^2\left(S_{xx}^{-1}(\overline{X}_1 - \overline{X}_0)^2 + n_1^{-1} + n_0^{-1}\right).\end{aligned} \tag{A.19}$$

We can now express the factor in the parenthesis with respect to a pooled variance estimate of the covariate included in the ANCOVA model, since a pooled variance estimate (as in equation (3.42)) can be expressed as

$$S_X = \sqrt{\frac{\sum_{w=0}^{1}\sum_{j=1}^{n_w}(X_{wj} - \overline{X}_w)^2}{n-2}}, \tag{A.20}$$

giving that $S_X^2(n-2) = S_{xx}$. Thus,

$$\mathbb{V}\mathrm{ar}\left(\widehat{\beta}_1 - \widehat{\beta}_0\right) = \sigma^2\left(\frac{(\overline{X}_1 - \overline{X}_0)^2}{S_X^2(n-2)} + n_1^{-1} + n_0^{-1}\right). \tag{A.21}$$

Using the usual unbiased estimator

$$\widehat{\sigma}^2 = \frac{(\mathbb{Y} - \mathbb{D}\widehat{\beta})^\top (\mathbb{Y} - \mathbb{D}\widehat{\beta})}{n-3} \sim \frac{\sigma^2}{n-3}\chi^2(n-3). \tag{A.22}$$

of σ, we obtain an unbiased estimator of the treatment effect variance as

$$\widehat{\mathbb{V}\mathrm{ar}}\left(\widehat{\beta}_1 - \widehat{\beta}_0\right) = \widehat{\sigma}^2 \left(\frac{(\overline{X}_1 - \overline{X}_0)^2}{S_X^2(n-2)} + n_1^{-1} + n_0^{-1} \right) \tag{A.23}$$

By inserting the expression for $\widehat{\beta}_X$ from equation (A.12) in equation (A.22) we can express $\widehat{\sigma}^2$ as

$$\begin{aligned}
&(n-3)\widehat{\sigma}^2 \\
&= \sum_{w=0}^{1}\sum_{j=1}^{n_w} \left(Y_{wj} - (\widehat{\beta}_w + X_{wj}\widehat{\beta}_X)\right)^2 \\
&= \sum_{w=0}^{1}\sum_{j=1}^{n_w} \left(Y_{wj} - (-\widehat{\beta}_X\overline{X}_w + \overline{Y}_w + X_{wj}\widehat{\beta}_X)\right)^2 \\
&= \sum_{w=0}^{1}\sum_{j=1}^{n_w} \left((Y_{wj} - \overline{Y}_w) - \widehat{\beta}_X(X_{wj} - \overline{X}_w)\right)^2 \\
&= \sum_{w=0}^{1}\sum_{j=1}^{n_w} \left[(Y_{wj} - \overline{Y}_w)^2 + \widehat{\beta}_X^2(X_{wj} - \overline{X}_w)^2 - 2\widehat{\beta}_X(Y_{wj} - \overline{Y}_w)(X_{wj} - \overline{X}_w)\right] \\
&= \sum_{w=0}^{1}\sum_{j=1}^{n_w}(Y_{wj} - \overline{Y}_w)^2 + \widehat{\beta}_X^2 \sum_{w=0}^{1}\sum_{j=1}^{n_w}(X_{wj} - \overline{X}_w)^2 - 2\widehat{\beta}_X \sum_{w=0}^{1}\sum_{j=1}^{n_w}(Y_{wj} - \overline{Y}_w)(X_{wj} - \overline{X}_w) \\
&= \sum_{w=0}^{1}\sum_{j=1}^{n_w}(Y_{wj} - \overline{Y}_w)^2 \qquad\qquad (A.24) \\
&\quad + \frac{\left[\sum_{w=0}^{1}\sum_{j=1}^{n_w}\left(X_{wj} - \overline{X}_w\right)\left(Y_{wj} - \overline{Y}_w\right)\right]^2}{\sum_{w=0}^{1}\sum_{j=1}^{n_w}(X_{wj} - \overline{X}_w)^2} - 2\frac{\left[\sum_{w=0}^{1}\sum_{j=1}^{n_w}\left(X_{wj} - \overline{X}_w\right)\left(Y_{wj} - \overline{Y}_w\right)\right]^2}{\sum_{w=0}^{1}\sum_{j=1}^{n_w}(X_{wj} - \overline{X}_w)^2} \\
&= \sum_{w=0}^{1}\sum_{j=1}^{n_w}(Y_{wj} - \overline{Y}_w)^2 + \widehat{\rho}^2 \sum_{w=0}^{1}\sum_{j=1}^{n_w}\left(Y_{wj} - \overline{Y}_w\right)^2 - 2\widehat{\rho}^2 \sum_{w=0}^{1}\sum_{j=1}^{n_w}\left(Y_{wj} - \overline{Y}_w\right)^2 \\
&= (1 - \widehat{\rho}^2)\sum_{w=0}^{1}\sum_{j=1}^{n_w}(Y_{wj} - \overline{Y}_w)^2,
\end{aligned}$$

where we have defined

$$\widehat{\rho} = \frac{\sum_{w=0}^{1}\sum_{j=1}^{n_w}(X_{wj} - \overline{X}_w)(Y_{wj} - \overline{Y}_w)}{\sqrt{\sum_{w=0}^{1}\sum_{j=1}^{n_w}(X_{wj} - \overline{X}_w)^2 \sum_{w=0}^{1}\sum_{j=1}^{n_w}(Y_{wj} - \overline{Y}_w)^2}} \tag{A.25}$$

as a pooled estimate of the correlation between X and Y.

Now, it follows from equations (A.19) and (A.24) that

$$\begin{aligned}\widehat{\mathbb{V}\mathrm{ar}}\left(\widehat{\beta}_1 - \widehat{\beta}_0\right) &= (1-\widehat{\rho}^2)\frac{\sum_{w=0}^{1}\sum_{j=1}^{n_w}(Y_{wj} - \overline{Y}_w)^2}{n-3}\left(S_{xx}^{-1}(\overline{X}_1 - \overline{X}_0)^2 + n_1^{-1} + n_0^{-1}\right) \\ &= (1-\widehat{\rho}^2)\frac{\sum_{w=0}^{1}\sum_{j=1}^{n_w}(Y_{wj} - \overline{Y}_w)^2}{n-3}\left(\frac{\left(\overline{X}_1 - \overline{X}_0\right)^2}{S_X^2(n-2)} + n_1^{-1} + n_0^{-1}\right).\end{aligned} \tag{A.26}$$

A.5 Positive Semidefiniteness of Covariance Matrix Difference

In example 2.3.1, where we investigate the effect of overspecification in the ANCOVA model on the estimator variances, we need the following lemma.

Lemma A.5.1.
For covariance matrices A and B, $B - A$ is positive semidefinite if and only if $A^{-1} - B^{-1}$ is positive semidefinite. ◀

Proof. We will show one implication, specifically that if $B - A$ is positive semidefinite, then $A^{-1} - B^{-1}$ is positive semidefinite, and the other implication then follows immediately.

Being covariance matrices, A and B are strictly positive definite and hence invertible with a well-defined square root. For a general positive definite matrix G, its inverse and square root are symmetric, so

$$\begin{aligned} G &= G^{1/2} I G^{1/2} \\ I &= G^{-1/2} G G^{-1/2}, \end{aligned} \tag{A.27}$$

where the latter expression is obtained by multiplying the left- and right hand sides with $G^{-1/2}$ from both sides.

For a general positive semidefinite matrix N, $C^\top N C$ is also positive semidefinite for C being any conformable matrix. From this fact, assuming that $B - A$ is positive semidefinite, using the second relation in (A.27) with A in place of G, and defining $M := A^{-1/2} B A^{-1/2}$, we get that

$$A^{-1/2}(B-A)A^{-1/2} = A^{-1/2}BA^{-1/2} - A^{-1/2}AA^{-1/2} = A^{-1/2}BA^{-1/2} - I = M - I \tag{A.28}$$

is also positive semidefinite. Using that M has an inverse and square root, due to being positive definite, we get that

$$M^{-1/2}(M-I)M^{-1/2} = I - M^{-1} = I - A^{1/2}B^{-1}A^{1/2} \tag{A.29}$$

is also positive semidefinite. Multiplying with $A^{-1/2}$ from both sides and using that the relations in equation (A.27) also hold for A^{-1}, we get that

$$\begin{aligned} A^{-1/2}\left(I - A^{1/2}B^{-1}A^{1/2}\right)A^{-1/2} &= A^{-1/2}IA^{-1/2} - A^{-1/2}A^{1/2}B^{-1}A^{1/2}A^{-1/2} \\ &= A^{-1} - B^{-1} \end{aligned} \tag{A.30}$$

is also positive semidefinite. ■

B FWER Bounds for Multiple Testing Procedures

In this appendix, we show for all presented multiple testing procedures in section 3.6, that the FWER is bounded.

B.1 Fixed Testing Sequence

Proposition B.1.1.
The FWER for the fixed testing sequence with each hypothesis tested at significance level α is bounded by α. ◀

Proof. We denote by s_i the number of non-rejected hypotheses for the first i hypotheses in the sequence. Then, we can regard α as a function of s_i, such that

$$\alpha(s_i) = \begin{cases} \alpha, & \text{if } s_i = 0 \\ 0, & \text{if } s_i > 0 \end{cases}. \tag{B.1}$$

To prove that the FWER is bounded by α for this procedure, we begin by assuming that we have m hypotheses that we want to test and that m_0 of these are true. The true hypotheses are denoted by $\mathcal{H}_0^1, \mathcal{H}_0^2, \ldots, \mathcal{H}_0^{m_0}$ with p-values $p_1, p_2, \ldots, p_{m_0}$. We denote the $m - m_0$ false hypotheses as $\mathcal{H}_0^{m_0+1}, \mathcal{H}_0^{m_0+2}, \ldots, \mathcal{H}_0^m$. According to the fixed testing sequence, we should order the hypotheses before testing. In the following, we will consider a specific ordering and a general ordering.

Let us first assume that we order the testing such that the $m - m_0$ false hypotheses are tested first. Furthermore we assume that the p-values $p_{m_0+1} = p_{m_0+2} = \cdots = p_m = 0$ for the false hypotheses $\mathcal{H}_0^{m_0+1}, \mathcal{H}_0^{m_0+2}, \ldots, \mathcal{H}_0^m$. This means that the false hypotheses are all correctly rejected, so that we do not make any type II errors. This configuration is now denoted L. In that case we have

$$\mathrm{FWER}_L = \mathbb{P}(p_1 < \alpha) = \alpha, \tag{B.2}$$

because if $p_1 \geqslant \alpha$ the procedure would stop, and no type I errors could occur; that is, the event $\{p_1 < \alpha\}$ contains all events in which at least one type I error occurs. We now wish to compare the FWER under this configuration with the FWER under a general configuration.

Let us now consider the events E_K and E_L that no false rejections (that is, no type I errors) are made under a general configuration K and the configuration L, respectively. Denote by t_i^K the

number of non-rejected hypotheses after testing i true hypotheses under configuration K, and equivalently for L. Then we have

$$\begin{aligned} E_K &= \bigcap_{i=1}^{m_0} \left\{ p_i \geqslant \alpha\left(t_{i-1}^K\right) \right\} \\ E_L &= \bigcap_{i=1}^{m_0} \left\{ p_i \geqslant \alpha\left(t_{i-1}^L\right) \right\} = \{p_1 \geqslant \alpha\}. \end{aligned} \tag{B.3}$$

When E_L occurs, we must have $t_i^L = i$ for $i = 1, 2, \ldots, m_0$; configuration L tests all the false hypotheses first, and we know that they are all rejected, so we know $t_0^L = 0$, and when testing the following m_0 true hypotheses, they must be non-rejected under E_L. When E_K occurs, we must have $t_i^K \geqslant i$ for $i = 1, 2, \ldots, m_0$; for configuration K, the false hypotheses potentially contribute to the number of non-rejected hypotheses, and all true hypotheses must still be non-rejected under E_K. Since $\alpha(t)$ is a decreasing function of t, we must then have $\alpha\left(t_i^K\right) \leqslant \alpha\left(t_i^L\right)$, so that $E_L \subseteq E_K$. Thus, we get the inequality in

$$\mathrm{FWER}_K = 1 - \mathbb{P}(E_K) \leqslant 1 - \mathbb{P}(E_L) = \mathrm{FWER}_L = \alpha. \tag{B.4}$$

Thus, for any configuration K of testing the m hypotheses in a fixed testing sequence, the probability of making a type I error is bounded by α [87]. ■

B.2 Bonferroni Corrections

Proposition B.2.1.
The FWER for the simple Bonferroni method is bounded by α. ◀

Proof. The Bonferroni correction method is based on Bonferroni's inequality, which states that for a countable set of events $A_1, A_2, A_3, \ldots$,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leqslant \sum_{i=1}^{\infty} \mathbb{P}(A_i). \tag{B.5}$$

If we again let $\mathcal{H}_0^1, \mathcal{H}_0^2, \ldots, \mathcal{H}_0^{m_0}$ be the true hypotheses, we obtain for the simple Bonferroni method that

$$\mathrm{FWER} = \mathbb{P}\left(\bigcup_{i=1}^{m_0} \{p_i < \alpha_i\}\right) \leqslant \sum_{i=1}^{m_0} \mathbb{P}\left(p_i < \alpha_i\right) = \sum_{i=1}^{m_0} \alpha_i \leqslant \alpha. \tag{B.6}$$

■

Proposition B.2.2.
The FWER for the Holm's sequentially rejective Bonferroni method is bounded by α. ◀

Proof. The Bonferroni-Holm procedure also ensures that the FWER is bounded by α, as seen by the following argument. Let $\mathcal{H}_0^{(k)}$ be the first wrongly rejected true hypothesis (that is, the first type I error). Then $\mathcal{H}_0^{(1)}, \mathcal{H}_0^{(2)}, \ldots, \mathcal{H}_0^{(k-1)}$ are correctly rejected false hypotheses. Using the same notation as in example 3.6.1, since $k-1 \leqslant m - m_0$, we have $\frac{1}{m-k+1} \leqslant \frac{1}{m_0}$. Furthermore, since $\mathcal{H}_0^{(k)}$ is rejected, we have

$$p_{(k)} < \frac{\alpha}{m-k+1} \leqslant \frac{\alpha}{m_0}. \tag{B.7}$$

If $\min_{i\in\{1,2,\ldots,m_0\}} p_i \geqslant \frac{\alpha}{m_0}$, we do not reject any of the true hypotheses. Thus,

$$\begin{aligned} \text{FWER} &\leqslant \mathbb{P}\left(\min_{i\in\{1,2,\ldots,m_0\}} p_i < \frac{\alpha}{m_0}\right) = \mathbb{P}\left(\bigcup_{i=1}^{m_0}\left\{p_i < \frac{\alpha}{m_0}\right\}\right) \\ &\leqslant \sum_{i=1}^{m_0} \mathbb{P}\left(p_i < \frac{\alpha}{m_0}\right) = \alpha, \end{aligned} \tag{B.8}$$

where the first inequality follows since the event of rejecting at least one true hypothesis is a subset of the event that $\min_{i\in\{1,2,\ldots,m_0\}} p_i \leqslant \frac{\alpha}{m_0}$. ■

C Theoretical Properties of Digital Twins

C.1 Efficient Influence Functions

C.1.1 Lemma 5.2.3

In the proof of lemma 5.2.3 we want to derive the Gâteaux derivative of the ATE. In this appendix we will derive equation (5.26) in further details. We will first inspect the part of the integrand in equation (5.25) that depends on t. Under regularity conditions, we are allowed to exchange the order of integration and the Gâteaux derivative. Moreover, we can use the product- and chain rule when taking the Gâteaux derivative. Hence, using shorthand notation and denoting by $f'|_{t=0}$ the Gâteaux derivative of f with respect to t, we get that

$$
\begin{aligned}
&\left(\frac{f_t(x,1,y)f_t(x)}{f_t(x,1)}\right)'\Bigg|_{t=0} \\
&= \left(\frac{\left(f_t(x,1,y)f_t(x)\right)' f_t(x,1) - f_t'(x,1)f_t(x,1,y)f_t(x)}{f_t(x,1)^2}\right)\Bigg|_{t=0} \\
&= \left(\frac{f_t'(x,1,y)f_t(x)f_t(x,1)}{f_t(x,1)^2} + \frac{f_t(x,1,y)f_t'(x)f_t(x,1)}{f_t(x,1)^2} - \frac{f_t'(x,1)f_t(x,1,y)f_t(x)}{f_t(x,1)^2}\right)\Bigg|_{t=0} \\
&= \left(\frac{f_t'(x,1,y)f_t(x)}{f_t(x,1)} + \frac{f_t(x,1,y)f_t'(x)}{f_t(x,1)} - \frac{f_t'(x,1)f_t(x,1,y)f_t(x)}{f_t(x,1)^2}\right)\Bigg|_{t=0} \\
&= \left(\frac{\left[\mathbb{1}_{x^*,w^*,y^*}(x,1,y) - f_t(x,1,y)\right] f_t(x)}{f_t(x,1)} + \frac{f_t(x,1,y)\left[\mathbb{1}_{x^*}(x) - f_t(x)\right]}{f_t(x,1)}\right. \\
&\quad \left. - \frac{\left[\mathbb{1}_{x^*,w^*}(x,1) - f_t(x,1)\right] f_t(x,1,y)f_t(x)}{f_t(x,1)^2}\right)\Bigg|_{t=0} \qquad \text{(C.1)} \\
&= \frac{\left[\mathbb{1}_{x^*,w^*,y^*}(x,1,y) - f(x,1,y)\right] f(x)}{f(x,1)} + \frac{f(x,1,y)\left[\mathbb{1}_{x^*}(x) - f(x)\right]}{f(x,1)} \\
&\quad - \frac{\left[\mathbb{1}_{x^*,w^*}(x,1) - f(x,1)\right] f(x,1,y)f(x)}{f(x,1)^2} \\
&= \frac{\mathbb{1}_{x^*,w^*,y^*}(x,1,y)f(x)}{f(x,1)} - \frac{f(x,1,y)f(x)}{f(x,1)} + \frac{\mathbb{1}_{x^*}(x)f(x,1,y)}{f(x,1)} - \frac{f(x)f(x,1,y)}{f(x,1)} \\
&\quad - \frac{\mathbb{1}_{x^*,w^*}(x,1)f(x,1,y)f(x)}{f(x,1)^2} + \frac{f(x,1)f(x,1,y)f(x)}{f(x,1)^2} \\
&= \frac{\mathbb{1}_{x^*,w^*,y^*}(x,1,y)f(x)}{f(x,1)} - \frac{\mathbb{1}_{x^*,w^*}(x,1)f(x,1,y)f(x)}{f(x,1)^2} + \frac{\mathbb{1}_{x^*}(x)f(x,1,y)}{f(x,1)} - \frac{f(x,1,y)f(x)}{f(x,1)},
\end{aligned}
$$

where the first equality follows from the chain- and product rules. The fourth equality follows from applying lemma 5.2.2 to all Gâteaux derivatives of f_t. Thus, when taking the Gâteaux derivative of (5.25), we get

$$\begin{aligned}\left.\frac{\mathrm{d}\Psi_1(F_t)}{\mathrm{d}t}\right|_{t=0} = &\int\int y\mathbb{1}_{x^*,w^*,y^*}(x,1,y)\frac{f(x)}{f(x,1)}\,\mathrm{d}y\,\mathrm{d}x - \int\int y\mathbb{1}_{x^*,w^*}(x,1)\frac{f(x,1,y)f(x)}{f(x,1)^2}\,\mathrm{d}y\,\mathrm{d}x \\ &+ \int\int \mathbb{1}_{x^*}(x)y\frac{f(x,1,y)}{f(x,1)}\,\mathrm{d}y\,\mathrm{d}x - \int\int y\frac{f(x,1,y)f(x)}{f(x,1)}\,\mathrm{d}y\,\mathrm{d}x. \qquad \text{(C.2)}\end{aligned}$$

We will now consider these four integrals one by one. From equation (5.25), we immediately recognise the last integral as $\Psi_1(F)$. The first integral can be expressed as

$$\begin{aligned}\int\int y\mathbb{1}_{x^*,w^*,y^*}(x,1,y)\frac{f(x)}{f(x,1)}\,\mathrm{d}y\,\mathrm{d}x &= \mathbb{1}_{w^*}(1)\int\int \mathbb{1}_{x^*,y^*}(x,y)y\frac{f(x)}{f(x,1)}\,\mathrm{d}y\,\mathrm{d}x \\ &= \mathbb{1}_{w^*}(1)\int\int \mathbb{1}_{x^*,y^*}(x,y)y\frac{1}{f(1\mid x)}\,\mathrm{d}y\,\mathrm{d}x \\ &= \mathbb{1}_{w^*}(1)\int \mathbb{1}_{y^*}(y)y\,\mathrm{d}y\int \mathbb{1}_{x^*}(x)\frac{1}{f\,(1\mid x)}\,\mathrm{d}x \qquad \text{(C.3)} \\ &= \mathbb{1}_{w^*}(1)\int y\,\mathrm{d}H_{y^*}(y)\int \frac{1}{f\,(1\mid x)}\,\mathrm{d}H_{x^*}(x) \\ &= \frac{\mathbb{1}_{w^*}(1)}{f\,(1\mid x^*)}y^*.\end{aligned}$$

The second integral in (C.2) can be expressed as

$$\begin{aligned}\int\int y\mathbb{1}_{x^*,w^*}(x,1)\frac{f(x,1,y)f(x)}{f(x,1)^2}\,\mathrm{d}y\,\mathrm{d}x &= \mathbb{1}_{w^*}(1)\int\int \mathbb{1}_{x^*}(x)y\frac{f(x,1,y)f(x)}{f(x,1)^2}\,\mathrm{d}y\,\mathrm{d}x \\ &= \mathbb{1}_{w^*}(1)\int\int y\frac{f(x,1,y)}{f(x,1)}\frac{f(x)}{f(x,1)}\,\mathrm{d}y\,\mathrm{d}H_{x^*}(x) \\ &= \mathbb{1}_{w^*}(1)\int\int y\frac{f(x,1,y)}{f(x,1)}\frac{1}{f(1\mid x)}\,\mathrm{d}y\,\mathrm{d}H_{x^*}(x) \\ &= \mathbb{1}_{w^*}(1)\int \frac{1}{f(1\mid x)}\int yf(y\mid x,1)\,\mathrm{d}y\,\mathrm{d}H_{x^*}(x) \qquad \text{(C.4)} \\ &= \frac{\mathbb{1}_{w^*}(1)}{f\,(1\mid x^*)}\int yf(y\mid x^*,1)\,\mathrm{d}y \\ &= \frac{\mathbb{1}_{w^*}(1)}{f\,(1\mid x^*)}\,\mathbb{E}_F\left[Y\mid X=x^*,W=1\right] \\ &= \frac{\mathbb{1}_{w^*}(1)}{f\,(1\mid x^*)}\gamma_1(x^*,F).\end{aligned}$$

For the third integral in (C.2), we get that

$$\begin{aligned}\int\int \mathbb{1}_{x^*}(x) y \frac{f(x,1,y)}{f(x,1)} \,\mathrm{d}y \,\mathrm{d}x &= \int\int \mathbb{1}_{x^*}(x) y f\left(y \mid x, 1\right) \,\mathrm{d}y \,\mathrm{d}x \\ &= \mathbb{E}_F\left[Y \mid X = x^*, W = 1\right] = \gamma_1(x^*, F).\end{aligned} \tag{C.5}$$

Using both the definition of the influence function in equation (5.14) as well as equations (C.2) to (C.5), we thus get

$$\begin{aligned}\left.\frac{\mathrm{d}\Psi_1(F_t)}{\mathrm{d}t}\right|_{t=0} &= \frac{\mathbb{1}_{w^*}(1)}{f\left(1 \mid x^*\right)} y^* - \frac{\mathbb{1}_{w^*}(1)}{f\left(1 \mid x^*\right)} \gamma_1(x^*, F) + \gamma_1(x^*, F) - \Psi_1(F) \\ &= \frac{\mathbb{1}_{w^*}(1)}{f\left(1 \mid x^*\right)} \left(y^* - \gamma_1(x^*, F)\right) + \gamma_1(x^*, F) - \Psi_1(F) = \varphi_1(x^*, w^*, y^*).\end{aligned} \tag{C.6}$$

C.2 Theorem 5.3.3

C.2.1 Inverse Matrix of $\mathbb{E}[D^\top D]$

We wish to determine the first factor $\mathbb{E}[D^\top D]^{-1}$ in equation (5.52). We begin by writing

$$D^\top D = \begin{bmatrix} 1 & W & X \\ W & W & WX \\ X^\top & X^\top W & X^\top X, \end{bmatrix}, \tag{C.7}$$

so

$$\mathbb{E}[D^\top D] = \begin{bmatrix} 1 & \mathbb{E}[W] & \mathbb{E}[X] \\ \mathbb{E}[W] & \mathbb{E}[W] & \mathbb{E}[W]\,\mathbb{E}[X] \\ \mathbb{E}[X^\top] & \mathbb{E}[X^\top]\,\mathbb{E}[W] & \mathbb{E}[X^\top X] \end{bmatrix}, \tag{C.8}$$

using that $\mathbb{C}\mathrm{ov}(W, X) = 0$ for an RCT. We propose a form of the inverse and check the validity of the inverse by multiplying by equation (C.8) and see that we get the identity. We propose that

$$\mathbb{E}\left[D^\top D\right]^{-1} = \begin{bmatrix} \frac{1}{1-\mathbb{E}[W]} + \mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{E}[X^\top] & -\frac{1}{1-\mathbb{E}[W]} & -\,\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}[X]^{-1} \\ -\frac{1}{1-\mathbb{E}[W]} & \frac{1}{\mathbb{E}[1-W]\,\mathbb{E}[W]} & 0 \\ -\left(\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\right)^\top & 0 & \mathbb{V}\mathrm{ar}(X)^{-1} \end{bmatrix}. \tag{C.9}$$

To see that this is truly the inverse of $\mathbb{E}[D^\top D]^{-1}$, we carry out calculations for each (block-)entry of $\mathbb{E}[D^\top D]\,\mathbb{E}[D^\top D]^{-1}$.

Entry $(1,1)$:

$$\frac{1}{1-\mathbb{E}[W]} + \mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{E}[X^\top] - \frac{\mathbb{E}[W]}{1-\mathbb{E}[W]} - \mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{E}[X^\top] = 1. \quad \text{(C.10)}$$

Entry $(1,2)$:

$$-\frac{1}{1-\mathbb{E}[W]} + \frac{\mathbb{E}[W]}{\mathbb{E}[1-W]\,\mathbb{E}[W]} = 0. \quad \text{(C.11)}$$

Block-entry $(1,3)$:

$$-\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1} + \mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1} = 0. \quad \text{(C.12)}$$

Entry $(2,1)$:

$$\frac{\mathbb{E}[W]}{1-\mathbb{E}[W]} + \mathbb{E}[W]\,\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{E}[X^\top] - \frac{\mathbb{E}[W]}{1-\mathbb{E}[W]} - \mathbb{E}[W]\,\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{E}[X^\top] = 0.$$

Entry $(2,2)$:

$$-\frac{\mathbb{E}[W]}{1-\mathbb{E}[W]} + \frac{\mathbb{E}[W]}{\mathbb{E}[1-W]\,\mathbb{E}[W]} = \frac{1}{\mathbb{E}[1-W]} - \frac{\mathbb{E}[W]}{\mathbb{E}[1-W]} = \frac{1-\mathbb{E}[W]}{\mathbb{E}[1-W]} = 1. \quad \text{(C.13)}$$

Block-entry $(2,3)$:

$$-\mathbb{E}[W]\,\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1} + \mathbb{E}[W]\,\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1} = 0. \quad \text{(C.14)}$$

Block-entry $(3,1)$:

$$\begin{aligned}
&\frac{\mathbb{E}[X^\top]}{1-\mathbb{E}[W]} + \mathbb{E}[X^\top]\,\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{E}[X^\top] - \frac{\mathbb{E}[X^\top]\,\mathbb{E}[W]}{1-\mathbb{E}[W]} - \mathbb{E}[X^\top X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{E}[X^\top] \\
&= \frac{\mathbb{E}[X^\top]}{1-\mathbb{E}[W]} + \mathbb{E}[X^\top]\,\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{E}[X^\top] - \frac{\mathbb{E}[X^\top]\,\mathbb{E}[W]}{1-\mathbb{E}[W]} \\
&\quad - \Big(\mathbb{V}\mathrm{ar}(X) + \mathbb{E}[X^\top]\,\mathbb{E}[X]\Big)\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{E}[X^\top] \qquad \text{(C.15)} \\
&= \frac{\mathbb{E}[X^\top]}{1-\mathbb{E}[W]} - \frac{\mathbb{E}[X^\top]\,\mathbb{E}[W]}{1-\mathbb{E}[W]} - \mathbb{E}[X^\top] \\
&= \frac{\mathbb{E}[X^\top](1-\mathbb{E}[W])}{1-\mathbb{E}[W]} - \mathbb{E}[X^\top] = 0.
\end{aligned}$$

Block-entry $(3,2)$:

$$-\frac{\mathbb{E}[X^\top]}{1-\mathbb{E}[W]}+\frac{\mathbb{E}[X^\top]\,\mathbb{E}[W]}{\mathbb{E}[1-W]\,\mathbb{E}[W]}=0. \tag{C.16}$$

Block-entry $(3,3)$:

$$\begin{aligned}
&-\mathbb{E}[X^\top]\,\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}+\mathbb{E}[X^\top X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\\
=\;&-\mathbb{E}[X^\top]\,\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}+\Big(\mathbb{V}\mathrm{ar}(X)+\mathbb{E}[X^\top]\,\mathbb{E}[X]\Big)\,\mathbb{V}\mathrm{ar}(X)^{-1}\\
=\;&I_p.
\end{aligned} \tag{C.17}$$

C.2.2 True parameters

We multiply the expressions (5.55) and (5.54) to get the true $\beta=\mathbb{E}\left[D^\top D\right]^{-1}\mathbb{E}[D^\top Y]$.

For the first entry of $\beta=(\beta_0,\beta_W,\beta_X)^\top$, we get

$$\begin{aligned}
\beta_0 &= \frac{\mathbb{E}[Y]}{1-\mathbb{E}[W]}+\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{E}[X^\top]\,\mathbb{E}[Y]-\frac{\mathbb{E}[W]\,\mathbb{E}[Y(1)]}{1-\mathbb{E}[W]}\\
&\quad-\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\Big(\mathbb{C}\mathrm{ov}(X^\top,Y)+\mathbb{E}[X^\top]\,\mathbb{E}[Y]\Big)\\
&= \frac{\mathbb{E}[Y]-\mathbb{E}[W]\,\mathbb{E}[Y(1)]}{1-\mathbb{E}[W]}-\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top,Y)\\
&= \frac{\mathbb{E}[Y-WY(1)]}{1-\mathbb{E}[W]}-\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top,Y)\\
&= \frac{\mathbb{E}[(1-W)Y(0)]}{1-\mathbb{E}[W]}-\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top,Y)\\
&= \mathbb{E}[Y(0)]-\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top,Y),
\end{aligned} \tag{C.18}$$

using in the fourth equality that $Y=WY(1)+(1-W)Y(0)$.

For the second entry, we get

$$\begin{aligned}
\beta_W &= -\frac{\mathbb{E}[Y]}{1-\mathbb{E}[W]}+\frac{\mathbb{E}[W]\,\mathbb{E}[Y(1)]}{\mathbb{E}[1-W]E[W]}\\
&= \frac{\mathbb{E}[Y(1)-Y]}{1-\mathbb{E}[W]}\\
&= \frac{\mathbb{E}[Y(1)-WY(1)-(1-W)Y(0)]}{1-\mathbb{E}[W]}\\
&= \frac{\mathbb{E}[(1-W)Y(1)-(1-W)Y(0)]}{1-\mathbb{E}[W]}\\
&= \mathbb{E}[Y(1)]-\mathbb{E}[Y(0)].
\end{aligned} \tag{C.19}$$

For the third "entry" (with dimension p being the same as that of X), we get

$$\begin{aligned}\beta_X &= -\left(\mathbb{E}[X]\,\mathbb{V}\mathrm{ar}(X)^{-1}\right)^\top \mathbb{E}[Y] + \mathbb{V}\mathrm{ar}(X)^{-1}\left(\mathbb{C}\mathrm{ov}(X^\top, Y) + \mathbb{E}[X^\top]\,\mathbb{E}[Y]\right) \\ &= -\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{E}[X^\top]\,\mathbb{E}[Y] + \mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top, Y) + \mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{E}[X^\top]\,\mathbb{E}[Y] \\ &= \mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}(X^\top, Y)\end{aligned} \tag{C.20}$$

C.3 Theorem 5.3.4

Having observations $D = (1, W, \widetilde{X}, W\widetilde{X})$, we can use the expression of $\beta = (\beta_0, \beta_W, \beta_X, \beta_{W\times X})^\top$ in equation (5.56), with $(\widetilde{X}, W\widetilde{X})$ playing the role of X. Specifically, we get

$$\beta = \begin{pmatrix} \mathbb{E}[Y(0)] \\ \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ \mathbb{V}\mathrm{ar}\left((\widetilde{X}, W\widetilde{X})\right)^{-1} \mathbb{C}\mathrm{ov}\left((\widetilde{X}, W\widetilde{X})^\top, Y\right) \end{pmatrix}, \tag{C.21}$$

using also that X is demeaned.

We now rewrite the last $2p$-dimensional entry of β. To start, we will derive the inverse matrix $\mathbb{V}\mathrm{ar}\left((\widetilde{X}, W\widetilde{X})\right)^{-1}$. One can verify that

$$\begin{bmatrix} \Upsilon\,\mathbb{V}\mathrm{ar}(W\widetilde{X}) & -\Upsilon\,\mathbb{C}\mathrm{ov}(\widetilde{X}, W\widetilde{X}) \\ -\Upsilon\,\mathbb{C}\mathrm{ov}(\widetilde{X}, W\widetilde{X})^\top & \Upsilon\,\mathbb{V}\mathrm{ar}(X) \end{bmatrix}, \tag{C.22}$$

where Υ, defined as

$$\Upsilon = \left(\mathbb{V}\mathrm{ar}(X)\,\mathbb{V}\mathrm{ar}(W\widetilde{X}) - \mathbb{C}\mathrm{ov}(\widetilde{X}, W\widetilde{X})\,\mathbb{C}\mathrm{ov}(\widetilde{X}, W\widetilde{X})^\top\right)^{-1} \tag{C.23}$$

is the inverse of

$$\mathbb{V}\mathrm{ar}\left((\widetilde{X}, W\widetilde{X})\right) = \begin{bmatrix} \mathbb{V}\mathrm{ar}(X) & \mathbb{C}\mathrm{ov}(\widetilde{X}, W\widetilde{X}) \\ \mathbb{C}\mathrm{ov}(\widetilde{X}, W\widetilde{X})^\top & \mathbb{V}\mathrm{ar}(W\widetilde{X}) \end{bmatrix}. \tag{C.24}$$

In the following, we will use the relations

$$\begin{aligned}\mathbb{V}\mathrm{ar}(W\widetilde{X}) &= \mathbb{C}\mathrm{ov}(W\widetilde{X}, W\widetilde{X}) = \mathbb{E}[W\widetilde{X}^\top\widetilde{X}] = \pi_1\,\mathbb{V}\mathrm{ar}(X) \\ \mathbb{C}\mathrm{ov}(\widetilde{X}, W\widetilde{X}) &= \mathbb{E}[\widetilde{X}^\top W\widetilde{X}] = \pi_1\,\mathbb{V}\mathrm{ar}(X) \\ \mathbb{C}\mathrm{ov}(W\widetilde{X}^\top, Y) &= \mathbb{E}[W\widetilde{X}^\top Y] = \mathbb{E}\left[W\widetilde{X}^\top\left(WY(1) + (1-W)Y(0)\right)\right] \\ &= \mathbb{E}[W\widetilde{X}^\top Y(1)] = \pi_1\,\mathbb{E}[\widetilde{X}Y(1)] = \pi_1\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}, Y(1)\right),\end{aligned} \tag{C.25}$$

which are obtained by using that X and W are independent, $W^2 = W$ and $W(1-W) = 0$. Furthermore, we will use that Υ can be expressed as

$$\begin{aligned}\Upsilon &= \left(\mathbb{V}\mathrm{ar}(X)\,\mathbb{V}\mathrm{ar}(W\widetilde{X}) - \mathbb{C}\mathrm{ov}(\widetilde{X}, W\widetilde{X})\,\mathbb{C}\mathrm{ov}(\widetilde{X}, W\widetilde{X})^\top\right)^{-1} \\ &= \left(\pi_1\,\mathbb{V}\mathrm{ar}(X)^2 - \pi_1^2\,\mathbb{V}\mathrm{ar}(X)^2\right)^{-1} = \left(\pi_1(1-\pi_1)\,\mathbb{V}\mathrm{ar}(X)^2\right)^{-1} \\ &= \left(\pi_1\pi_0\,\mathbb{V}\mathrm{ar}(X)^2\right)^{-1} = \frac{1}{\pi_1\pi_0}\,\mathbb{V}\mathrm{ar}(X)^{-2}.\end{aligned} \tag{C.26}$$

Using these relations and the expression of the inverse in equation (C.22), we find that the last $2p$-dimensional entry of β can be expressed as

$$\begin{aligned}&\mathbb{V}\mathrm{ar}\left((\widetilde{X}, W\widetilde{X})\right)^{-1}\mathbb{C}\mathrm{ov}\left((\widetilde{X}, W\widetilde{X})^\top, Y\right) \\ &= \begin{bmatrix}\Upsilon\,\mathbb{V}\mathrm{ar}(W\widetilde{X}) & -\Upsilon\,\mathbb{C}\mathrm{ov}(\widetilde{X}, W\widetilde{X}) \\ -\Upsilon\,\mathbb{C}\mathrm{ov}(\widetilde{X}, W\widetilde{X})^\top & \Upsilon\,\mathbb{V}\mathrm{ar}(X),\end{bmatrix}\begin{bmatrix}\mathbb{C}\mathrm{ov}(\widetilde{X}^\top, Y) \\ \mathbb{C}\mathrm{ov}(W\widetilde{X}^\top, Y)\end{bmatrix} \\ &= \begin{bmatrix}\Upsilon\left(\mathbb{V}\mathrm{ar}(W\widetilde{X})\,\mathbb{C}\mathrm{ov}(\widetilde{X}^\top, Y) - \mathbb{C}\mathrm{ov}(\widetilde{X}, W\widetilde{X})\,\mathbb{C}\mathrm{ov}(W\widetilde{X}^\top, Y)\right) \\ -\Upsilon\left(\mathbb{C}\mathrm{ov}(\widetilde{X}, W\widetilde{X})^\top\,\mathbb{C}\mathrm{ov}(\widetilde{X}^\top, Y) - \mathbb{V}\mathrm{ar}(X)\,\mathbb{C}\mathrm{ov}(W\widetilde{X}^\top, Y)\right)\end{bmatrix} \\ &= \begin{bmatrix}\frac{1}{\pi_0\pi_1}\,\mathbb{V}\mathrm{ar}(X)^{-2}\left(\mathbb{V}\mathrm{ar}(W\widetilde{X})\,\mathbb{C}\mathrm{ov}(\widetilde{X}^\top, Y) - \mathbb{C}\mathrm{ov}(\widetilde{X}, W\widetilde{X})\,\mathbb{C}\mathrm{ov}(W\widetilde{X}^\top, Y)\right) \\ \frac{1}{\pi_0\pi_1}\,\mathbb{V}\mathrm{ar}(X)^{-2}\left(-\mathbb{C}\mathrm{ov}(\widetilde{X}, W\widetilde{X})^\top\,\mathbb{C}\mathrm{ov}(\widetilde{X}^\top, Y) + \mathbb{V}\mathrm{ar}(X)\,\mathbb{C}\mathrm{ov}(W\widetilde{X}^\top, Y)\right)\end{bmatrix} \\ &= \begin{bmatrix}\frac{1}{\pi_0\pi_1}\,\mathbb{V}\mathrm{ar}(X)^{-2}\left(\pi_1\,\mathbb{V}\mathrm{ar}(X)\,\mathbb{C}\mathrm{ov}(\widetilde{X}^\top, Y) - \pi_1^2\,\mathbb{V}\mathrm{ar}(X)\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(1)\right)\right) \\ \frac{1}{\pi_0\pi_1}\,\mathbb{V}\mathrm{ar}(X)^{-2}\left(-\pi_1\,\mathbb{V}\mathrm{ar}(X)\,\mathbb{C}\mathrm{ov}(\widetilde{X}^\top, Y) + \pi_1\,\mathbb{V}\mathrm{ar}(X)\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(1)\right)\right)\end{bmatrix} \\ &= \begin{bmatrix}\frac{1}{\pi_0}\,\mathbb{V}\mathrm{ar}(X)^{-1}\left(\mathbb{C}\mathrm{ov}(\widetilde{X}^\top, Y) - \pi_1\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(1)\right)\right) \\ \frac{1}{\pi_0}\,\mathbb{V}\mathrm{ar}(X)^{-1}\left(-\mathbb{C}\mathrm{ov}(\widetilde{X}^\top, Y) + \mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(1)\right)\right)\end{bmatrix}.\end{aligned} \tag{C.27}$$

That is, $\beta = (\beta_0, \beta_W, \beta_X, \beta_{W\times X})^\top$ can be expressed as

$$\beta = \begin{pmatrix}\mathbb{E}[Y(0)] \\ \text{ATE} \\ \frac{1}{\pi_0}\,\mathbb{V}\mathrm{ar}(X)^{-1}\left(\mathbb{C}\mathrm{ov}(\widetilde{X}^\top, Y) - \pi_1\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(1)\right)\right) \\ \frac{1}{\pi_0}\,\mathbb{V}\mathrm{ar}(X)^{-1}\left(-\mathbb{C}\mathrm{ov}(\widetilde{X}^\top, Y) + \mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(1)\right)\right)\end{pmatrix} \tag{C.28}$$

From this expression, we get

$$\begin{aligned}\gamma_w(\widetilde{X}, F_{II}) =& \left(1, w, \widetilde{X}, w\widetilde{X}\right)\beta \\ =& \mathbb{E}[Y(0)] + w \cdot \text{ATE} \\ &+ \widetilde{X}\frac{1}{\pi_0}\mathbb{V}\text{ar}(X)^{-1}\left(\mathbb{C}\text{ov}(\widetilde{X}^\top, Y) - \pi_1\,\mathbb{C}\text{ov}\left(\widetilde{X}^\top, Y(1)\right)\right) \\ &+ w \cdot \widetilde{X}\frac{1}{\pi_0}\mathbb{V}\text{ar}(X)^{-1}\left(-\mathbb{C}\text{ov}(\widetilde{X}^\top, Y) + \mathbb{C}\text{ov}\left(\widetilde{X}^\top, Y(1)\right)\right),\end{aligned} \tag{C.29}$$

so that

$$\begin{aligned}\gamma_1(\widetilde{X}, F_{II}) =& \mathbb{E}[Y(0)] + \text{ATE} \\ &+ \widetilde{X}\frac{1}{\pi_0}\mathbb{V}\text{ar}(X)^{-1}\left(-\pi_1\,\mathbb{C}\text{ov}\left(\widetilde{X}^\top, Y(1)\right) + \mathbb{C}\text{ov}\left(\widetilde{X}^\top, Y(1)\right)\right) \\ =& \mathbb{E}[Y(1)] + \widetilde{X}\frac{1}{\pi_0}(1-\pi_1)\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}\left(\widetilde{X}^\top, Y(1)\right) \\ =& \mathbb{E}[Y(1)] + \widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}\left(\widetilde{X}^\top, Y(1)\right),\end{aligned} \tag{C.30}$$

and

$$\begin{aligned}\gamma_0(\widetilde{X}, F_{II}) =& \mathbb{E}[Y(0)] + \widetilde{X}\frac{1}{\pi_0}\mathbb{V}\text{ar}(X)^{-1}\left(\mathbb{C}\text{ov}(\widetilde{X}^\top, Y) - \pi_1\,\mathbb{C}\text{ov}\left(\widetilde{X}^\top, Y(1)\right)\right) \\ =& \mathbb{E}[Y(0)] + \widetilde{X}\frac{1}{\pi_0}\mathbb{V}\text{ar}(X)^{-1}\left(\pi_0\,\mathbb{C}\text{ov}\left(\widetilde{X}^\top, Y(0)\right)\right) \\ =& \mathbb{E}[Y(0)] + \widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}\left(\widetilde{X}^\top, Y(0)\right),\end{aligned} \tag{C.31}$$

where we use the relation in equation (5.72) in the second equality. By taking the expectation with respect to $\widetilde{X}$ in the above expressions, we obtain $\Psi_w(F_{II}) = \mathbb{E}[Y(w)]$. Using this with lemma 5.2.3, we get

$$\begin{aligned}\varphi_1 =& \frac{W}{\pi_1}\left[Y - \left(\mathbb{E}[Y(1)] + \widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}\left(\widetilde{X}^\top, Y(1)\right)\right)\right] \\ &+ \left(\mathbb{E}[Y(1)] + \widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}\left(\widetilde{X}^\top, Y(1)\right)\right) - \mathbb{E}[Y(1)] \\ =& \frac{W}{\pi_1}\left[Y - \left(\mathbb{E}[Y(1)] + \widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}\left(\widetilde{X}^\top, Y(1)\right)\right)\right] \\ &+ \widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}\left(\widetilde{X}^\top, Y(1)\right) \\ =& \frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1)]\right) - \frac{W-\pi_1}{\pi_1}\widetilde{X}\,\mathbb{V}\text{ar}(X)^{-1}\,\mathbb{C}\text{ov}\left(\widetilde{X}^\top, Y(1)\right),\end{aligned} \tag{C.32}$$

and

$$\begin{aligned}
\varphi_0 &= \frac{1-W}{\pi_0}\left[Y - \left(\mathbb{E}[Y(0)] + \widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(0)\right)\right)\right] \\
&\quad + \mathbb{E}[Y(0)] + \widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(0)\right) - \mathbb{E}[Y(0)] \\
&= \frac{1-W}{\pi_0}\left[Y - \mathbb{E}[Y(0)] - \widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(0)\right)\right] \\
&\quad + \widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(0)\right) \\
&= \frac{1-W}{\pi_0}\left(Y - \mathbb{E}[Y(0)]\right) - \frac{(1-W)-\pi_0}{\pi_0}\widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(0)\right),
\end{aligned} \tag{C.33}$$

so that

$$\begin{aligned}
\varphi_{\mathrm{ATE},II} &= \varphi_1 - \varphi_0 \\
&= \frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1)]\right) - \frac{1-W}{\pi_0}\left(Y - \mathbb{E}[Y(0)]\right) \\
&\quad - \frac{W-\pi_1}{\pi_1}\widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(1)\right) \\
&\quad + \frac{(1-W)-\pi_0}{\pi_0}\widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(0)\right) \\
&= \frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1)]\right) - \frac{1-W}{\pi_0}\left(Y - \mathbb{E}[Y(0)]\right) \\
&\quad - \left(\frac{W-\pi_1}{\pi_1}\widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(1)\right)\right. \\
&\quad \left. + \frac{W-\pi_1}{\pi_0}\widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(0)\right)\right) \\
&= \frac{W}{\pi_1}\left(Y - \mathbb{E}[Y(1)]\right) - \frac{1-W}{\pi_0}\left(Y - \mathbb{E}[Y(0)]\right) \\
&\quad - \frac{W-\pi_1}{\pi_0\pi_1}\widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\left(\pi_0\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(1)\right) + \pi_1\,\mathbb{C}\mathrm{ov}\left(\widetilde{X}^\top, Y(0)\right)\right) \\
&= \varphi_\Delta - \frac{W-\pi_1}{\pi_0\pi_1}\widetilde{X}\,\mathbb{V}\mathrm{ar}(X)^{-1}\xi_*^\top.
\end{aligned} \tag{C.34}$$

C.4 Corollary 5.3.5

In this appendix, we wish to derive an expression of the difference between the asymptotic variances in equations (5.83) and (5.84). We first obtain

$$\begin{aligned}\begin{bmatrix}\Sigma_X & \zeta^\top \\ \zeta & \sigma_M^2\end{bmatrix}^{-1} &= \begin{bmatrix}\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1} & -\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1}\zeta^\top\sigma_M^{-2} \\ -\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\zeta\Sigma_X^{-1} & \left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\end{bmatrix} \\ &= \begin{bmatrix}\Sigma_X^{-1} + \Sigma_X^{-1}\zeta^\top\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\zeta\Sigma_X^{-1} & -\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1}\zeta^\top\sigma_M^{-2} \\ -\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\zeta\Sigma_X^{-1} & \sigma_M^{-2} + \sigma_M^{-2}\zeta\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1}\zeta^\top\sigma_M^{-2}\end{bmatrix}.\end{aligned} \tag{C.35}$$

In each of the diagonal (block-)entries in the second expression we use the Woodbury matrix identity to obtain the third expression. In the following derivations, we will use the second expression, but one can more easily verify that this is in fact the inverse using the third expression, as we will do entry-wise in the following.

Block-entry $(1,1)$:

$$\Sigma_X\left(\Sigma_X^{-1} + \Sigma_X^{-1}\zeta^\top\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\zeta\Sigma_X^{-1}\right) + \zeta^\top\left(-\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\zeta\Sigma_X^{-1}\right) = I_p$$

Block-entry $(2,1)$:

$$\begin{aligned}&\zeta\left(\Sigma_X^{-1} + \Sigma_X^{-1}\zeta^\top\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\zeta\Sigma_X^{-1}\right) + \sigma_M^2\left(-\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\zeta\Sigma_X^{-1}\right) \\ &= \left(1 + \zeta\Sigma_X^{-1}\zeta^\top\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1} - \sigma_M^2\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\right)\zeta\Sigma_X^{-1} \\ &= \left(1 - \left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\right)\zeta\Sigma_X^{-1} = 0\end{aligned}$$

Block-entry $(1,2)$:

$$\begin{aligned}&\Sigma_X\left(-\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1}\zeta^\top\sigma_M^{-2}\right) + \zeta^\top\left(\sigma_M^{-2} + \sigma_M^{-2}\zeta\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1}\zeta^\top\sigma_M^{-2}\right) \\ &= \left(I_p + \zeta^\top\sigma_M^{-2}\zeta\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1} - \Sigma_X\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1}\right)\zeta^\top\sigma_M^{-2} \\ &= \left(I_p - \left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1}\right)\zeta^\top\sigma_M^{-2} = 0\end{aligned}$$

Entry $(2,2)$:

$$\zeta\left(-\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1}\zeta^\top\sigma_M^{-2}\right) + \sigma_M^2\left(\sigma_M^{-2} + \sigma_M^{-2}\zeta\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1}\zeta^\top\sigma_M^{-2}\right) = 1$$

Using the second expression in (C.35) of the inverse covariance matrix in (5.84), we can now subtract the asymptotic variance expression in (5.84) from the one in (5.83), in order to obtain

$$
\begin{aligned}
&\frac{1}{\pi_0\pi_1}\Bigg(- \xi_{X*}\Sigma_X^{-1}\xi_{X*}^\top \\
&+ \begin{bmatrix}\xi_{X*} & \xi_{M*}\end{bmatrix}\begin{bmatrix}\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1} & -\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1}\zeta^\top\sigma_M^{-2} \\ -\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\zeta\Sigma_X^{-1} & \left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\end{bmatrix}\begin{bmatrix}\xi_{X*}^\top \\ \xi_{M*}\end{bmatrix}\Bigg) \\
&= \frac{1}{\pi_0\pi_1}\left(\begin{bmatrix}\xi_{X*} & \xi_{M*}\end{bmatrix}\begin{bmatrix}\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1}\xi_{X*}^\top - \left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1}\zeta^\top\sigma_M^{-2}\xi_{M*} \\ -\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\zeta\Sigma_X^{-1}\xi_{X*}^\top + \left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\xi_{M*}\end{bmatrix} - \xi_{X*}\Sigma_X^{-1}\xi_{X*}^\top\right) \\
&= \frac{1}{\pi_0\pi_1}\Bigg(\xi_{X*}\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1}\xi_{X*}^\top - \xi_{M*}\xi_{X*}\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1}\zeta^\top\sigma_M^{-2} \qquad \text{(C.36)} \\
&\quad - \xi_{M*}\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\zeta\Sigma_X^{-1}\xi_{X*}^\top + \xi_{M*}^2\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1} - \xi_{X*}\Sigma_X^{-1}\xi_{X*}^\top\Bigg) \\
&= \frac{1}{\pi_0\pi_1}\Bigg(\xi_{X*}\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1}\xi_{X*}^\top - \xi_{M*}\xi_{X*}\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1}\zeta^\top\sigma_M^{-2} \\
&\quad + \frac{\xi_{M*}\left(\xi_{M*} - \zeta\Sigma_X^{-1}\xi_{X*}^\top\right)}{\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top} - \xi_{X*}\Sigma_X^{-1}\xi_{X*}^\top\Bigg).
\end{aligned}
$$

Again using the Woodbury identity as we did in (C.35), we get that

$$
\left(\Sigma_X - \zeta^\top\sigma_M^{-2}\zeta\right)^{-1} = \Sigma_X^{-1} + \Sigma_X^{-1}\zeta^\top\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\zeta\Sigma_X^{-1}. \tag{C.37}
$$

Inserting this in (C.36), we get the difference between the asymptotic variance of the estimator using X and the asymptotic variance of the estimator using (X, M) by first defining the scalars

$$
\begin{aligned}
a &= \xi_{X*}\Sigma_X^{-1}\zeta^\top = \zeta\Sigma^{-1}\xi_{X*}^\top \\
b &= \zeta\Sigma_X^{-1}\zeta^\top.
\end{aligned} \tag{C.38}
$$

Then

$$
\begin{aligned}
&\frac{1}{\pi_0\pi_1}\Bigg[\xi_{X*}\Sigma_X^{-1}\xi_{X*}^\top + \xi_{X*}\Sigma_X^{-1}\zeta^\top\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\zeta\Sigma_X^{-1}\xi_{X*}^\top \\
&- \left(\xi_{M*}\xi_{X*}\Sigma_X^{-1}\zeta^\top\sigma_M^{-2} + \xi_{M*}\xi_{X*}\Sigma_X^{-1}\zeta^\top\left(\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top\right)^{-1}\zeta\Sigma_X^{-1}\zeta^\top\sigma_M^{-2}\right) \\
&- \frac{\xi_{M*}\left(\xi_{M*} - \zeta\Sigma^{-1}\xi_{X*}^\top\right)}{\sigma_M^2 - \zeta\Sigma_X^{-1}\zeta^\top} - \xi_{X*}\Sigma_X^{-1}\xi_{X*}^\top\Bigg] \\
&= \frac{1}{\pi_0\pi_1}\left[a(\sigma_M^2 - b)^{-1}a - \left(\xi_{M*}a\sigma_M^{-2} + \xi_{M*}a(\sigma_M^2 - b)^{-1}b\sigma_M^{-2}\right) + \frac{\xi_{M*}^2 - \xi_{M*}a}{\sigma_M^2 - b}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\pi_0\pi_1}\left[\frac{a^2+\xi_{M*}^2-\xi_{M*}a}{\sigma_M^2-b} - \xi_{M*}a\left(\sigma_M^{-2}+(\sigma_M^2-b)^{-1}b\sigma_M^{-2}\right)\right] \\
&= \frac{1}{\pi_0\pi_1}\left[\frac{a^2+\xi_{M*}^2-\xi_{M*}a}{\sigma_M^2-b} - \xi_{M*}a\left(\frac{\sigma_M^{-2}(\sigma_M^2-b)+b\sigma_M^{-2}}{\sigma_M^2-b}\right)\right] \\
&= \frac{1}{\pi_0\pi_1}\left[\frac{a^2+\xi_{M*}^2-\xi_{M*}a}{\sigma_M^2-b} - \xi_{M*}a\frac{1}{\sigma_M^2-b}\right] \\
&= \frac{1}{\pi_0\pi_1}\frac{a^2+\xi_{M*}^2-2\xi_{M*}a}{\sigma_M^2-b} \\
&= \frac{1}{\pi_0\pi_1}\frac{(\xi_{M*}-a)^2}{\sigma_M^2-b} \\
&= \frac{1}{\pi_0\pi_1}\frac{\left(\xi_{M*}-\xi_{X*}\Sigma_X^{-1}\zeta^\top\right)^2}{\sigma_M^2-\zeta\Sigma_X^{-1}\zeta^\top}
\end{aligned}
\tag{C.39}
$$

C.5 Lemma 5.3.7

Consider using $f(X) = \big(\mathbb{E}[Y(0)\,|\,X],\, \mathbb{E}[Y(1)\,|\,X]\big)$ in equation (5.91). Define $R_w(X) = Y(w) - \mathbb{E}[Y(w)\,|\,X]$ for $w = 0, 1$ which has mean 0 by the law of total expectation. For any function $g\colon \mathcal{X} \to \mathbb{R}^k$ (with row vector output), we have

$$
\begin{aligned}
\mathbb{C}\mathrm{ov}\left(R_w(X), g(X)\right) &= \mathbb{E}\left[\left(R_w(X) - \mathbb{E}[R_w(X)]\right)\left(g(X) - \mathbb{E}\left[g(X)\right]\right)\right] \\
&= \mathbb{E}\left[\left(Y(w) - \mathbb{E}[Y(w)\,|\,X]\right)\left(g(X) - \mathbb{E}\left[g(X)\right]\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(Y(w) - \mathbb{E}[Y(w)\,|\,X]\right)\left(g(X) - \mathbb{E}\left[g(X)\right]\right)\,\middle|\,X\right]\right] \\
&= \mathbb{E}\left[\left(\mathbb{E}[Y(w)\,|\,X] - \mathbb{E}[Y(w)\,|\,X]\right)\left(g(X) - \mathbb{E}\left[g(X)\right]\right)\right] \\
&= 0,
\end{aligned}
\tag{C.40}
$$

where we use the law of total expectation in the third equality, and in the fourth equality we use that given X, $g(X)$ is deterministic, so that we can use linearity of the expected value. Then we have

$$
\begin{aligned}
\mathbb{C}\mathrm{ov}\left(Y(w),\, \mathbb{E}[Y(w)\,|\,X]\right) &= \mathbb{C}\mathrm{ov}\left(R_w(X) + \mathbb{E}[Y(w)\,|\,X],\, \mathbb{E}[Y(w)\,|\,X]\right) \\
&= \mathbb{C}\mathrm{ov}\left(R_w(X),\, \mathbb{E}[Y(w)\,|\,X]\right) + \mathbb{V}\mathrm{ar}\left(\mathbb{E}[Y(w)\,|\,X]\right) \\
&= \mathbb{V}\mathrm{ar}\left(\mathbb{E}[Y(w)\,|\,X]\right)
\end{aligned}
\tag{C.41}
$$

and

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(0)\,|\,X]\right) &= \mathbb{C}\mathrm{ov}\left(R_1(X) + \mathbb{E}[Y(1)\,|\,X], \mathbb{E}[Y(0)\,|\,X]\right) \\ &= \mathbb{C}\mathrm{ov}\left(R_1(X),\, \mathbb{E}[Y(0)\,|\,X]\right) + \mathbb{C}\mathrm{ov}\left(\mathbb{E}[Y(1)\,|\,X],\, \mathbb{E}[Y(0)\,|\,X]\right) \\ &= \mathbb{C}\mathrm{ov}\left(\mathbb{E}[Y(1)\,|\,X],\, \mathbb{E}[Y(0)\,|\,X]\right), \end{aligned} \tag{C.42}$$

as well as

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left(Y(0),\, \mathbb{E}[Y(1)\,|\,X]\right) &= \mathbb{C}\mathrm{ov}\left(R_0(X) + \mathbb{E}[Y(0)\,|\,X], \mathbb{E}[Y(1)\,|\,X]\right) \\ &= \mathbb{C}\mathrm{ov}\left(R_0(X),\, \mathbb{E}[Y(1)\,|\,X]\right) + \mathbb{C}\mathrm{ov}\left(\mathbb{E}[Y(0)\,|\,X],\, \mathbb{E}[Y(1)\,|\,X]\right) \\ &= \mathbb{C}\mathrm{ov}\left(\mathbb{E}[Y(0)\,|\,X],\, \mathbb{E}[Y(1)\,|\,X]\right). \end{aligned} \tag{C.43}$$

The last two equations imply that $\mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(0)\,|\,X]\right) = \mathbb{C}\mathrm{ov}\left(Y(0),\, \mathbb{E}[Y(1)\,|\,X]\right)$. Inserting these expressions into $\mathbb{V}\mathrm{ar}\left(f(X)\right)^{-1}$ and $\xi_{f*}^{\top}$, we obtain

$$\begin{aligned} &\mathbb{V}\mathrm{ar}\left(f(X)\right)^{-1} \\ &= \begin{bmatrix} \mathbb{V}\mathrm{ar}\left(\mathbb{E}[Y(0)\,|\,X]\right) & \mathbb{C}\mathrm{ov}\left(\mathbb{E}[Y(0)\,|\,X],\, \mathbb{E}[Y(1)\,|\,X]\right) \\ \mathbb{C}\mathrm{ov}\left(\mathbb{E}[Y(0)\,|\,X],\, \mathbb{E}[Y(1)\,|\,X]\right) & \mathbb{V}\mathrm{ar}\left(\mathbb{E}[Y(1)\,|\,X]\right) \end{bmatrix}^{-1} \\ &= \frac{1}{\mathbb{V}\mathrm{ar}\left(\mathbb{E}[Y(0)\,|\,X]\right)\mathbb{V}\mathrm{ar}\left(\mathbb{E}[Y(1)\,|\,X]\right) - \mathbb{C}\mathrm{ov}\left(\mathbb{E}[Y(0)\,|\,X],\, \mathbb{E}[Y(1)\,|\,X]\right)^2} \\ &\quad \cdot \begin{bmatrix} \mathbb{V}\mathrm{ar}\left(\mathbb{E}[Y(1)\,|\,X]\right) & -\mathbb{C}\mathrm{ov}\left(\mathbb{E}[Y(0)\,|\,X],\, \mathbb{E}[Y(1)\,|\,X]\right) \\ -\mathbb{C}\mathrm{ov}\left(\mathbb{E}[Y(0)\,|\,X],\, \mathbb{E}[Y(1)\,|\,X]\right) & \mathbb{V}\mathrm{ar}\left(\mathbb{E}[Y(0)\,|\,X]\right) \end{bmatrix} \\ &= \frac{1}{\mathbb{C}\mathrm{ov}\left(Y(0),\, \mathbb{E}[Y(0)\,|\,X]\right)\mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(1)\,|\,X]\right) - \mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(0)\,|\,X]\right)^2} \\ &\quad \cdot \begin{bmatrix} \mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(1)\,|\,X]\right) & -\mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(0)\,|\,X]\right) \\ -\mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(0)\,|\,X]\right) & \mathbb{C}\mathrm{ov}\left(Y(0),\, \mathbb{E}[Y(0)\,|\,X]\right) \end{bmatrix}, \end{aligned} \tag{C.44}$$

and

$$\begin{aligned} \xi_{f*}^{\top} &= \pi_0\, \mathbb{C}\mathrm{ov}\left(Y(1), \left(\mathbb{E}[Y(0)\,|\,X],\ \mathbb{E}[Y(1)\,|\,X]\right)\right)^{\top} \\ &\qquad + \pi_1\, \mathbb{C}\mathrm{ov}\left(Y(0), \left(\mathbb{E}[Y(0)\,|\,X],\ \mathbb{E}[Y(1)\,|\,X]\right)\right)^{\top} \\ &= \pi_0 \begin{pmatrix} \mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(0)\,|\,X]\right) \\ \mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(1)\,|\,X]\right) \end{pmatrix} + \pi_1 \begin{pmatrix} \mathbb{C}\mathrm{ov}\left(Y(0),\, \mathbb{E}[Y(0)\,|\,X]\right) \\ \mathbb{C}\mathrm{ov}\left(Y(0),\, \mathbb{E}[Y(1)\,|\,X]\right) \end{pmatrix} \\ &= \pi_0 \begin{pmatrix} \mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(0)\,|\,X]\right) \\ \mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(1)\,|\,X]\right) \end{pmatrix} + \pi_1 \begin{pmatrix} \mathbb{C}\mathrm{ov}\left(Y(0),\, \mathbb{E}[Y(0)\,|\,X]\right) \\ \mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(0)\,|\,X]\right) \end{pmatrix}. \end{aligned} \tag{C.45}$$

Combining these, we obtain

$$
\begin{aligned}
&\mathbb{V}\mathrm{ar}\left(f(X)\right)^{-1}\xi_{f*}^{\top} \\
&= \frac{1}{\mathbb{C}\mathrm{ov}\left(Y(0),\, \mathbb{E}[Y(0)\,|\,X]\right)\mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(1)\,|\,X]\right) - \mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(0)\,|\,X]\right)^2} \\
&\quad \cdot \left(\pi_0 \begin{bmatrix} 0 \\ \mathbb{C}\mathrm{ov}\left(Y(0),\, \mathbb{E}[Y(0)\,|\,X]\right)\mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(1)\,|\,X]\right) - \mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(0)\,|\,X]\right)^2 \end{bmatrix} \right. \\
&\quad \left. + \pi_1 \begin{bmatrix} \mathbb{C}\mathrm{ov}\left(Y(0),\, \mathbb{E}[Y(0)\,|\,X]\right)\mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(1)\,|\,X]\right) - \mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(0)\,|\,X]\right)^2 \\ 0 \end{bmatrix} \right) \\
&= \begin{pmatrix} \pi_1 \\ \pi_0 \end{pmatrix}. \qquad \text{(C.46)}
\end{aligned}
$$

C.6 Corollary 5.3.8

We note that since $\mathbb{E}[Y(1)\,|\,X] = \mathbb{E}[Y(0)\,|\,X] + \mathrm{ATE}$, we have that

$$\mathbb{C}\mathrm{ov}\left(\mathbb{E}[Y(1)\,|\,X],\, \mathbb{E}[Y(0)\,|\,X]\right) = \mathbb{V}\mathrm{ar}\left(\mathbb{E}[Y(0)\,|\,X]\right) = \mathbb{V}\mathrm{ar}\left(\mathbb{E}[Y(1)\,|\,X]\right). \qquad \text{(C.47)}$$

Inserting the expression of $f(X) = \mathbb{E}[Y(0)\,|\,X]$ in (5.78) in place of X in the expression of $\mathbb{V}\mathrm{ar}(X)^{-1}$, we obtain

$$\mathbb{V}\mathrm{ar}\left(f(X)\right)^{-1} = \mathbb{V}\mathrm{ar}\left(\mathbb{E}[Y(0)\,|\,X]\right)^{-1}, \qquad \text{(C.48)}$$

and similarly by inserting $f(X) = \mathbb{E}[Y(0)\,|\,X]$ in place of X in the expression of ξ_{f*}, we obtain

$$
\begin{aligned}
\xi_{f*} &= \pi_0\,\mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(0)\,|\,X]\right) + \pi_1\,\mathbb{C}\mathrm{ov}\left(Y(0),\, \mathbb{E}[Y(0)\,|\,X]\right) \\
&= (\pi_0 + \pi_1)\,\mathbb{C}\mathrm{ov}\left(Y(1),\, \mathbb{E}[Y(0)\,|\,X]\right) \\
&= \mathbb{C}\mathrm{ov}\left(\mathbb{E}[Y(1)\,|\,X],\, \mathbb{E}[Y(0)\,|\,X]\right) \\
&= \mathbb{V}\mathrm{ar}\left(\mathbb{E}[Y(0)\,|\,X]\right),
\end{aligned} \qquad \text{(C.49)}
$$

where we use equation (5.80) in the second equality with $f(X) = \mathbb{E}[Y(0)\,|\,X]$ in place of X to obtain a common factor, in the third equality we use (C.42) and in the last equality we use equation (C.47). Combining these expressions, we obtain

$$\mathbb{V}\mathrm{ar}\left(f(X)\right)^{-1}\xi_{f*} = 1. \qquad \text{(C.50)}$$

C.7 Lemma 5.3.10

In order to prove lemma 5.3.10, we will first need to prove the following two lemmas.

Lemma C.7.1.
Let $f\colon \mathcal{X} \to \mathbb{R}$ be a bounded function on a compact set $\mathcal{X}$, $f_n\colon \mathcal{X} \to \mathbb{R}$ be a sequence of uniformly bounded random functions such that $|f(X) - f_n(X)| \xrightarrow{L^2} 0$, where $X \in \mathcal{X}$ is a random variable independent of f_n. Then

$$\begin{aligned} \mathbb{E}_X[f_n(X)] &\xrightarrow{\mathbb{P}} \mathbb{E}[f(X)] \\ \mathbb{C}\mathrm{ov}_X\left(f(X), f_n(X)\right) &\xrightarrow{\mathbb{P}} \mathbb{V}\mathrm{ar}\left(f(X)\right) \\ \mathbb{V}\mathrm{ar}\left(f_n(X)\right) &\xrightarrow{\mathbb{P}} \mathbb{V}\mathrm{ar}\left(f(X)\right). \end{aligned} \tag{C.51}$$

◀

Proof. Denoting by $\mathbb{P}_{f_n}$ the probability measure of f_n and $\mathbb{P}_X$ the probability measure of X, the probability measure of (f_n, X) factorises from independence to $\mathbb{P} = \mathbb{P}_{f_n}\,\mathbb{P}_X$. We then get

$$\begin{aligned} \mathbb{E}_{f_n}\left[\left(\mathbb{E}_X[f_n(X)] - \mathbb{E}[f(X)]\right)^2\right] &= \int \left(\mathbb{E}_X[f_n(X)] - \mathbb{E}[f(X)]\right)^2 \,\mathrm{d}\,\mathbb{P}_{f_n} \\ &= \int \left(\int f_n(X) - f(X)\,\mathrm{d}\,\mathbb{P}_X\right)^2 \,\mathrm{d}\,\mathbb{P}_{f_n} \\ &\leqslant \int\int \left(f_n(X) - f(X)\right)^2 \,\mathrm{d}\,\mathbb{P}_X \,\mathrm{d}\,\mathbb{P}_{f_n} \\ &= \int \left(f_n(X) - f(X)\right)^2 \,\mathrm{d}\left(\mathbb{P}_X \otimes \mathbb{P}_{f_n}\right) \\ &= \int \left(f_n(X) - f(X)\right)^2 \,\mathrm{d}\,\mathbb{P} \\ &\longrightarrow 0, \end{aligned} \tag{C.52}$$

where the inequality follows from Jensen's inequality. The second to last equality holds by Fubini's theorem, using that a probability measure space is σ-finite and that the integrand is measurable with respect to the product measure $\mathbb{P}_X \otimes \mathbb{P}_{f_n}$ by the convergence assumption that $|f(X) - f_n(X)| \xrightarrow{L^2} 0$, so that the integrand with respect to $\mathbb{P}_X \otimes \mathbb{P}_{f_n}$ must be defined. In the last equality, we use that the product measure of two independent random variables is the product of the measures. The convergence holds using the assumption that $|f(X) - f_n(X)| \xrightarrow{L^2} 0$. This implies that $\mathbb{E}_X[f_n(X)] \xrightarrow{L^2} \mathbb{E}[f(X)]$ and hence this convergence also holds in probability.

Now we will show that $\mathbb{E}_X[f(X)f_n(X)] \xrightarrow{\mathbb{P}} \mathbb{E}[f(X)^2]$. Using similar arguments as above, we

obtain

$$\begin{aligned}\mathbb{E}_{f_n}\left[\left(\mathbb{E}_X[f(X)f_n(X)] - \mathbb{E}[f(X)^2]\right)^2\right] &\leqslant \int \Big(f(X)\left(f_n(X) - f(X)\right)\Big)^2 \,\mathrm{d}\,\mathbb{P} \\ &\leqslant K^2 \int \left(f_n(X) - f(X)\right)^2 \,\mathrm{d}\,\mathbb{P} \\ &\longrightarrow 0,\end{aligned} \tag{C.53}$$

where we have assumed that f is bounded by K. Again using similar arguments leads to $\mathbb{E}_X[f_n(X)^2] \xrightarrow{\mathbb{P}} \mathbb{E}[f(X)^2]$ if we assume that the bound K is also a uniform bound for f_n. Specifically,

$$\begin{aligned}\mathbb{E}_{f_n}\left[\left(\mathbb{E}_X[f_n(X)^2] - \mathbb{E}[f(X)^2]\right)^2\right] &\leqslant \int \left(f_n(X)^2 - f(X)^2\right)^2 \,\mathrm{d}\,\mathbb{P} \\ &= \int \Big(\left(f_n(X) + f(X)\right)\left(f_n(X) - f(X)\right)\Big)^2 \,\mathrm{d}\,\mathbb{P} \\ &\leqslant 4K^2 \int \left(f_n(X) - f(X)\right)^2 \,\mathrm{d}\,\mathbb{P} \\ &\longrightarrow 0.\end{aligned} \tag{C.54}$$

We can now use the above results to obtain

$$\begin{aligned}\mathbb{C}\mathrm{ov}_X\left(f(X), f_n(X)\right) &= \mathbb{E}_X[f(X)f_n(X)] - \mathbb{E}[f(X)]\,\mathbb{E}_X[f_n(X)] \\ &\xrightarrow{\mathbb{P}} \mathbb{E}[f(X)^2] - \mathbb{E}[f(X)]^2 = \mathbb{V}\mathrm{ar}\left(f(X)\right).\end{aligned} \tag{C.55}$$

Furthermore, we have

$$\begin{aligned}\mathbb{V}\mathrm{ar}_X\left(f_n(X)\right) &= \mathbb{E}_X[f_n(X)^2] - \mathbb{E}[f_n(X)]^2 \\ &\xrightarrow{\mathbb{P}} \mathbb{E}_X[f(X)^2] - \mathbb{E}[f(X)]^2 = \mathbb{V}\mathrm{ar}\left(f(X)\right),\end{aligned} \tag{C.56}$$

using Slutsky's theorem (the probability limit of a product of two sequences of random variables with constant probability limits is the product of these limits) for the convergence of the second term. ■

Lemma C.7.2.
Let $f\colon \mathcal{X} \to \mathbb{R}$ be a bounded function on a compact set $\mathcal{X}$, $f_n\colon \mathcal{X} \to \mathbb{R}$ be a sequence of uniformly bounded random functions such that $|f(X) - f_n(X)| \xrightarrow{L^2} 0$, where $X \in \mathcal{X}$ is a random variable independent of f_n. Assume furthermore that $\mathbb{V}\mathrm{ar}_X\left(f_n(X)\right) > \varepsilon$ for all $n \in \mathbb{N}$ and some fixed $\varepsilon > 0$. Then, $(1 - B_n) \xrightarrow{L^2} 0$ for

$$B_n = \frac{\mathbb{C}\mathrm{ov}_X\left(f(X), f_n(X)\right)}{\mathbb{V}\mathrm{ar}_X\left(f_n(X)\right)}. \tag{C.57}$$

◀

Proof. We wish to show that $(1 - B_n) \xrightarrow{L^2} 0$. This corresponds to showing that

$$\forall \omega \; \exists N \in \mathbb{N} : \quad n \geqslant N \quad \Rightarrow \quad \mathbb{E}\left[(1 - B_n)^2\right] < \omega. \tag{C.58}$$

First, we note that since f_n and f are (uniformly) bounded by some $K > 0$, we have for all n that

$$\begin{aligned} \left|\mathbb{C}\mathrm{ov}\left(f_n(X), f(X)\right)\right| &= \left|\mathbb{E}[f_n(X)f(X)] - \mathbb{E}[f_n(X)]\,\mathbb{E}[f(X)]\right| \\ &\leqslant \left|\int f_n(X)f(X)\,\mathrm{d}\,\mathbb{P}\right| + \left|\int f_n(X)\,\mathrm{d}\,\mathbb{P}\int f(X)\,\mathrm{d}\,\mathbb{P}\right| \\ &\leqslant 2K^2. \end{aligned} \tag{C.59}$$

Hence, B_n is uniformly bounded by $\frac{2K^2}{\varepsilon}$. From lemma C.7.1 and Slutsky's theorem, we have the convergence $B_n \xrightarrow{\mathbb{P}} 1$, which means that

$$\forall \delta, \gamma > 0 \; \exists N \in \mathbb{N} : \quad n \geqslant N \quad \Rightarrow \quad \mathbb{P}\left(|1 - B_n| \geqslant \delta\right) < \gamma. \tag{C.60}$$

We are now ready to show (C.58). For arbitrarily small ω, we know from (C.60) that specifically for $\delta(\omega) := \sqrt{\frac{\omega}{2}}$ and $\gamma(\omega) := \frac{\omega}{2\left(1+\frac{2K^2}{\varepsilon}\right)^2}$, we can find an N' such that $\mathbb{P}\left(|1 - B_n| \geqslant \delta(\omega)\right) < \gamma(\omega)$. We now obtain for an arbitrarily small ω that

$$\begin{aligned} \mathbb{E}\left[(1 - B_n)^2\right] &= \int (1 - B_n)^2\,\mathrm{d}\,\mathbb{P} \\ &= \int_{|1-B_n|<\delta(\omega)} (1 - B_n)^2\,\mathrm{d}\,\mathbb{P} + \int_{|1-B_n|\geqslant\delta(\omega)} (1 - B_n)^2\,\mathrm{d}\,\mathbb{P} \\ &\leqslant \delta(\omega)^2 + \left(1 + \frac{2K^2}{\varepsilon}\right)^2 \mathbb{P}\left(|1 - B_n| \geqslant \delta(\omega)\right) \\ &< \frac{\omega}{2} + \frac{\omega}{2} = \omega, \end{aligned} \tag{C.61}$$

for all $n \geqslant N'$. We conclude that $(1 - B_n) \xrightarrow{L^2} 0$. ■

Now, we are able to restate and prove lemma 5.3.10 on the next page.

Lemma 5.3.10.
Let $f\colon \mathcal{X} \to \mathbb{R}$ be a bounded function on a compact set $\mathcal{X}$, $f_n\colon \mathcal{X} \to \mathbb{R}$ be a sequence of uniformly bounded random functions such that $|f(X) - f_n(X)| \xrightarrow{L^2} 0$, where $X \in \mathcal{X}$ is a random variable independent of f_n. Assume furthermore that $\mathbb{V}\mathrm{ar}_X\left(f_n(X)\right) > \varepsilon$ for all $n \in \mathbb{N}$ and some fixed $\varepsilon > 0$. Then

$$|f(X) - f_n(X)B_n| \xrightarrow{L^2} 0 \tag{C.62}$$

for

$$B_n = \frac{\mathbb{C}\mathrm{ov}_X\left(f(X), f_n(X)\right)}{\mathbb{V}\mathrm{ar}_X\left(f_n(X)\right)}. \tag{C.63}$$

◀

Proof. By the triangle inequality, and f_n being uniformly bounded by K, we have almost surely that

$$\begin{aligned} |f(x) - f_n(x)B_n| &\leqslant |f(x) - f_n(x)| + |f_n(x) - f_n(x)B_n| \\ &\leqslant |f(x) - f_n(x)| + K\,|1 - B_n|\,, \end{aligned} \tag{C.64}$$

so that, almost surely,

$$\left(f(x) - f_n(x)B_n\right)^2 \leqslant \left(f(x) - f_n(x)\right)^2 + K^2\left(1 - B_n\right)^2 + 2K\,|f(x) - f_n(x)| \cdot |1 - B_n|\,.$$

Using this, the result now follows from

$$\begin{aligned} \mathbb{E}\left[\left(f(X) - f_n(X)B_n\right)^2\right] &\leqslant \mathbb{E}\left[\left(f(X) - f_n(X)\right)^2\right] + K^2\,\mathbb{E}\left[(1 - B_n)^2\right] \\ &\quad + 2K\,\mathbb{E}\left[|f(x) - f_n(x)| \cdot |1 - B_n|\right] \\ &\leqslant \mathbb{E}\left[\left(f(X) - f_n(X)\right)^2\right] + K^2\,\mathbb{E}\left[(1 - B_n)^2\right] \\ &\quad + 4K^2\,\mathbb{E}\left[|1 - B_n|\right] \\ &\longrightarrow 0, \end{aligned} \tag{C.65}$$

where the second inequality follows from K being a uniform bound on f_n and f, so that, almost surely, $|f(x) - f_n(x)|$ is bounded by $2K$. The convergence of each of the terms follows from $(1 - B_n) \xrightarrow{L^2} 0$ (and hence also $(1 - B_n) \xrightarrow{L^1} 0$), shown in lemma C.7.2, and the assumption that $|f(X) - f_n(X)| \xrightarrow{L^2} 0$. ■

D Comparison of Approaches

D.1 Performance in Different Scenarios

(Additional) results from the simulation study in section 6.2, displayed in figure 6.2, are listed in table D.1.

D.2 Overspecification and Underspecification

(Additional) results from the simulation study in section 6.3, displayed in figure 6.5, are listed in table D.2.

Scenario	Covariates	Interaction	Model	Estimate	SD	RMSE	Power	Coverage	Type I error
linear	No	No	None	3.01	0.56	0.56	0.96	0.95	0.03
linear	No	No	PSM	3.00	0.50	0.33	1.00	1.00	0.00
linear	No	No	Random	3.01	0.56	0.56	0.96	0.95	0.03
linear	No	No	RF	3.00	0.12	0.12	1.00	0.95	0.03
linear	No	No	Oracle0	3.00	0.09	0.09	1.00	0.96	0.02
linear	No	Yes	None	3.01	0.56	0.56	0.96	0.95	0.03
linear	No	Yes	PSM	3.00	0.50	0.33	1.00	1.00	0.00
linear	No	Yes	Random	3.01	0.56	0.56	0.96	0.95	0.03
linear	No	Yes	RF	3.00	0.12	0.12	1.00	0.94	0.03
linear	No	Yes	Oracle0	3.00	0.09	0.09	1.00	0.96	0.02
linear	Yes	No	None	3.00	0.09	0.09	1.00	0.96	0.02
linear	Yes	No	PSM	3.00	0.08	0.10	1.00	0.90	0.05
linear	Yes	No	Random	3.00	0.09	0.09	1.00	0.96	0.02
linear	Yes	No	RF	3.00	0.09	0.09	1.00	0.96	0.02
linear	Yes	No	Oracle0	3.00	0.09	0.09	1.00	0.96	0.02
linear	Yes	Yes	None	3.00	0.09	0.09	1.00	0.96	0.02
linear	Yes	Yes	PSM	3.00	0.08	0.10	1.00	0.91	0.04
linear	Yes	Yes	Random	3.00	0.09	0.09	1.00	0.96	0.03
linear	Yes	Yes	RF	3.00	0.09	0.09	1.00	0.96	0.02
linear	Yes	Yes	Oracle0	3.00	0.09	0.09	1.00	0.96	0.02
homogeneous	No	No	None	3.05	1.43	1.48	0.31	0.94	0.03
homogeneous	No	No	PSM	2.98	1.47	1.52	0.30	0.94	0.03
homogeneous	No	No	Random	3.05	1.43	1.49	0.31	0.94	0.04
homogeneous	No	No	RF	3.00	0.36	0.37	1.00	0.94	0.02
homogeneous	No	No	Oracle0	3.01	0.09	0.09	1.00	0.95	0.03
homogeneous	No	Yes	None	3.05	1.43	1.48	0.31	0.94	0.03
homogeneous	No	Yes	PSM	2.98	1.47	1.52	0.29	0.94	0.03
homogeneous	No	Yes	Random	3.05	1.43	1.48	0.31	0.94	0.04
homogeneous	No	Yes	RF	3.01	0.36	0.37	1.00	0.94	0.03
homogeneous	No	Yes	Oracle0	3.01	0.09	0.09	1.00	0.95	0.02
homogeneous	Yes	No	None	3.06	1.33	1.36	0.36	0.94	0.03
homogeneous	Yes	No	PSM	3.01	1.39	1.49	0.34	0.93	0.04
homogeneous	Yes	No	Random	3.06	1.33	1.36	0.35	0.95	0.03
homogeneous	Yes	No	RF	2.99	0.35	0.35	1.00	0.94	0.03
homogeneous	Yes	No	Oracle0	3.01	0.09	0.09	1.00	0.96	0.03
homogeneous	Yes	Yes	None	3.10	1.33	1.37	0.37	0.94	0.03
homogeneous	Yes	Yes	PSM	3.02	1.38	1.51	0.34	0.93	0.04
homogeneous	Yes	Yes	Random	3.10	1.33	1.36	0.37	0.94	0.03
homogeneous	Yes	Yes	RF	2.99	0.35	0.36	1.00	0.93	0.03
homogeneous	Yes	Yes	Oracle0	3.01	0.09	0.09	1.00	0.96	0.03

(continued on next page)

(continued from previous page)

heterogeneous	No	No	None	3.03	1.61	1.52	0.23	0.97	0.01
heterogeneous	No	No	PSM	2.96	1.59	1.51	0.23	0.96	0.01
heterogeneous	No	No	Random	3.04	1.61	1.52	0.23	0.96	0.02
heterogeneous	No	No	RF	2.99	0.50	0.48	0.99	0.96	0.02
heterogeneous	No	No	Oracle0	2.99	0.42	0.36	1.00	0.98	0.01
heterogeneous	No	Yes	None	3.03	1.61	1.52	0.23	0.97	0.01
heterogeneous	No	Yes	PSM	2.96	1.59	1.51	0.23	0.96	0.01
heterogeneous	No	Yes	Random	3.04	1.61	1.51	0.23	0.96	0.01
heterogeneous	No	Yes	RF	3.00	0.48	0.47	0.99	0.96	0.02
heterogeneous	No	Yes	Oracle0	3.00	0.41	0.36	1.00	0.97	0.01
heterogeneous	Yes	No	None	3.05	1.35	1.36	0.34	0.95	0.02
heterogeneous	Yes	No	PSM	3.01	1.40	1.49	0.32	0.93	0.03
heterogeneous	Yes	No	Random	3.05	1.35	1.37	0.34	0.95	0.02
heterogeneous	Yes	No	RF	3.00	0.45	0.46	0.99	0.95	0.03
heterogeneous	Yes	No	Oracle0	3.00	0.29	0.31	1.00	0.93	0.03
heterogeneous	Yes	Yes	None	3.09	1.32	1.36	0.38	0.94	0.02
heterogeneous	Yes	Yes	PSM	3.01	1.38	1.47	0.33	0.93	0.03
heterogeneous	Yes	Yes	Random	3.09	1.32	1.36	0.37	0.94	0.03
heterogeneous	Yes	Yes	RF	2.99	0.35	0.45	1.00	0.87	0.06
heterogeneous	Yes	Yes	Oracle0	3.00	0.09	0.28	1.00	0.47	0.26
covariate shift	No	No	None	2.99	1.61	1.49	0.22	0.97	0.02
covariate shift	No	No	PSM	1.13	1.65	2.27	0.01	0.85	0.00
covariate shift	No	No	Random	2.98	1.61	1.49	0.22	0.97	0.02
covariate shift	No	No	RF	3.01	0.88	0.90	0.62	0.95	0.03
covariate shift	No	No	Oracle0	2.99	0.42	0.35	1.00	0.98	0.01
covariate shift	No	Yes	None	2.99	1.61	1.49	0.22	0.97	0.02
covariate shift	No	Yes	PSM	1.14	1.65	2.27	0.01	0.85	0.00
covariate shift	No	Yes	Random	2.99	1.61	1.49	0.22	0.97	0.02
covariate shift	No	Yes	RF	3.04	0.85	0.90	0.66	0.94	0.04
covariate shift	No	Yes	Oracle0	3.00	0.41	0.35	1.00	0.98	0.02
covariate shift	Yes	No	None	3.00	1.35	1.32	0.31	0.97	0.02
covariate shift	Yes	No	PSM	4.26	1.41	1.75	0.66	0.88	0.12
covariate shift	Yes	No	Random	2.99	1.35	1.32	0.31	0.97	0.02
covariate shift	Yes	No	RF	3.02	0.76	0.76	0.76	0.96	0.03
covariate shift	Yes	No	Oracle0	3.01	0.29	0.31	1.00	0.94	0.04
covariate shift	Yes	Yes	None	3.03	1.32	1.32	0.34	0.96	0.02
covariate shift	Yes	Yes	PSM	3.54	1.36	1.35	0.47	0.96	0.04
covariate shift	Yes	Yes	Random	3.03	1.32	1.32	0.34	0.96	0.02
covariate shift	Yes	Yes	RF	3.02	0.70	0.76	0.80	0.94	0.03
covariate shift	Yes	Yes	Oracle0	3.01	0.09	0.28	1.00	0.46	0.27

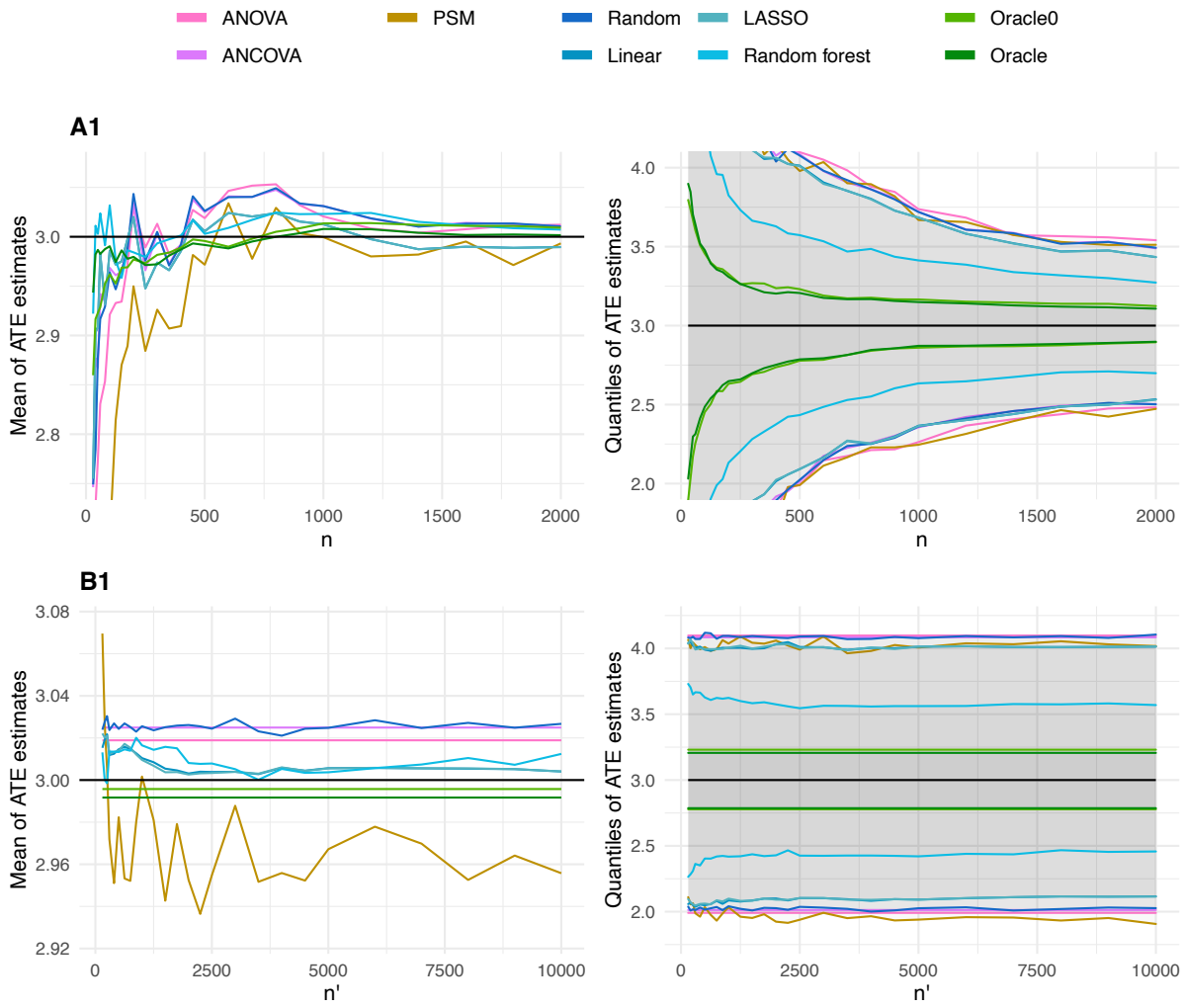
Table D.1: Empirical means of AN(C)OVA model ATE estimates and estimated standard errors, RMSE and empirically estimated power, coverage and type I error rate using $1,000$ simulated data sets under all scenarios presented in table 6.1. "covariates" column indicates whether raw covariate adjustments were included in the AN(C)OVA model for all 10 simulated covariates. "interaction" column indicates whether interaction terms between treatment allocation and all raw covariates (as well as the estimated prognostic score for models "Random", "RF" and "Oracle0") were included in the AN(C)OVA model. RMSE and std.err are displayed in figure 6.2.

Scenario	Covariates	Model	Estimate	SD	RMSE	Power	Coverage	Type I error
linear	No	None	3.01	0.56	0.56	0.96	0.95	0.03
linear	No	PSM	2.99	0.50	0.39	0.99	0.99	0.01
linear	No	Random	3.01	0.56	0.56	0.95	0.95	0.03
linear	No	RF	3.00	0.23	0.24	1.00	0.94	0.03
linear	No	Oracle0	3.00	0.09	0.09	1.00	0.96	0.02
linear	Yes	None	3.00	0.38	0.38	1.00	0.94	0.03
linear	Yes	PSM	3.00	0.34	0.31	1.00	0.96	0.02
linear	Yes	Random	3.00	0.38	0.38	1.00	0.94	0.03
linear	Yes	RF	3.00	0.23	0.24	1.00	0.95	0.02
linear	Yes	Oracle0	3.00	0.09	0.09	1.00	0.96	0.02
homogeneous	No	None	3.05	1.43	1.48	0.31	0.94	0.03
homogeneous	No	PSM	3.00	1.48	1.51	0.29	0.95	0.03
homogeneous	No	Random	3.05	1.43	1.48	0.32	0.93	0.04
homogeneous	No	RF	3.02	0.75	0.77	0.75	0.94	0.03
homogeneous	No	Oracle0	3.01	0.09	0.09	1.00	0.95	0.03
homogeneous	Yes	None	3.04	1.37	1.41	0.32	0.94	0.04
homogeneous	Yes	PSM	2.99	1.43	1.48	0.32	0.95	0.03
homogeneous	Yes	Random	3.03	1.37	1.41	0.32	0.93	0.04
homogeneous	Yes	RF	3.03	0.75	0.76	0.76	0.94	0.03
homogeneous	Yes	Oracle0	3.01	0.09	0.09	1.00	0.95	0.03
heterogeneous	No	None	3.03	1.61	1.52	0.23	0.97	0.01
heterogeneous	No	PSM	2.97	1.60	1.50	0.24	0.96	0.01
heterogeneous	No	Random	3.04	1.61	1.52	0.23	0.97	0.01
heterogeneous	No	RF	3.00	0.87	0.82	0.63	0.96	0.02
heterogeneous	No	Oracle0	2.99	0.42	0.36	1.00	0.98	0.01
heterogeneous	Yes	None	3.04	1.46	1.44	0.30	0.96	0.02
heterogeneous	Yes	PSM	2.99	1.49	1.49	0.29	0.95	0.02
heterogeneous	Yes	Random	3.04	1.47	1.44	0.30	0.96	0.02
heterogeneous	Yes	RF	3.01	0.86	0.82	0.65	0.96	0.02
heterogeneous	Yes	Oracle0	2.99	0.36	0.34	1.00	0.96	0.02
covariate shift	No	None	2.99	1.61	1.49	0.22	0.97	0.02
covariate shift	No	PSM	1.33	2.08	2.20	0.01	0.95	0.00
covariate shift	No	Random	2.98	1.61	1.49	0.22	0.97	0.02
covariate shift	No	RF	3.01	1.06	1.07	0.47	0.95	0.03
covariate shift	No	Oracle0	2.99	0.42	0.35	1.00	0.98	0.01
covariate shift	Yes	None	2.98	1.46	1.40	0.27	0.96	0.02
covariate shift	Yes	PSM	3.19	1.92	1.40	0.14	0.99	0.01
covariate shift	Yes	Random	2.98	1.46	1.40	0.27	0.96	0.02
covariate shift	Yes	RF	3.02	1.04	1.05	0.49	0.94	0.03
covariate shift	Yes	Oracle0	3.00	0.36	0.33	1.00	0.97	0.02

Table D.2: Empirical means of AN(C)OVA model ATE estimates and estimated standard errors, RMSE and empirically estimated power, coverage and type I error rate using $1,000$ simulated data sets under all scenarios presented in table 6.1. "covariates" column indicates whether raw covariate adjustments were included in the AN(C)OVA model for all 10 simulated covariates. No interaction terms were included in the AN(C)OVA models. RMSE and std.err are displayed in figure 6.5.

D.3 Varying Sample Sizes

Results of simulations varying only n and only n', respectively, are displayed in figure D.1. Figures A1-A4 display results of simulations where only the size of the current RCT arm n is varied, while figures B1-B4 display results of simulations where only the number of historical data points n' is varied.



(continued on next page)

(continued from previous page)

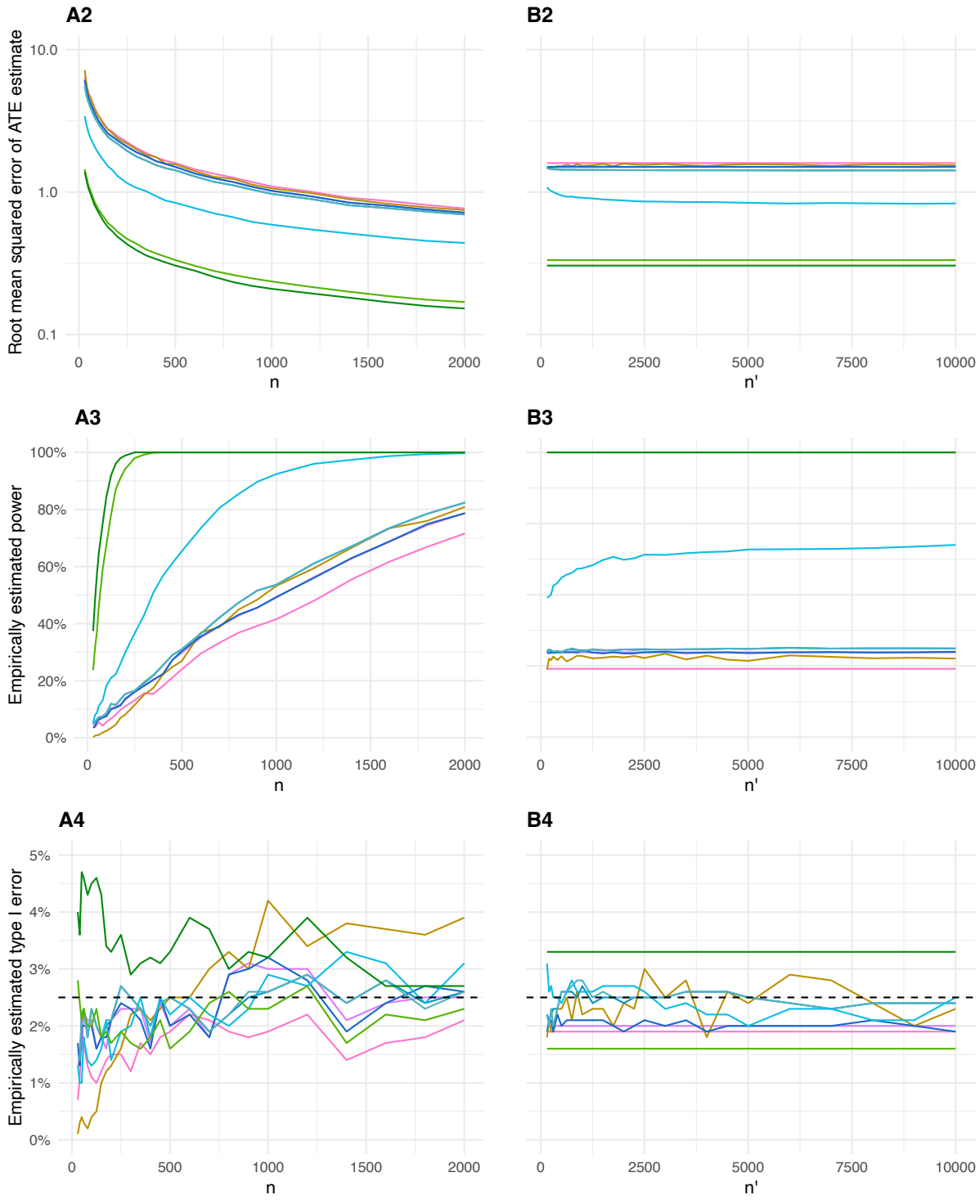


Figure D.1: Performance measures for all investigated models. For fixed $n' = 5,000$ varying n (A1–A4) and fixed $n = 500$ varying n' (B1–B4). Descriptions of all plots are available in figures 6.6–6.8.

D.4 Prospective Power Estimation in Homogeneous Case

Figure D.2 displays prospective power estimations in the scenario of a homogeneous treatment effect, as opposed to the heterogeneous scenario presented in chapter 6.4.2.

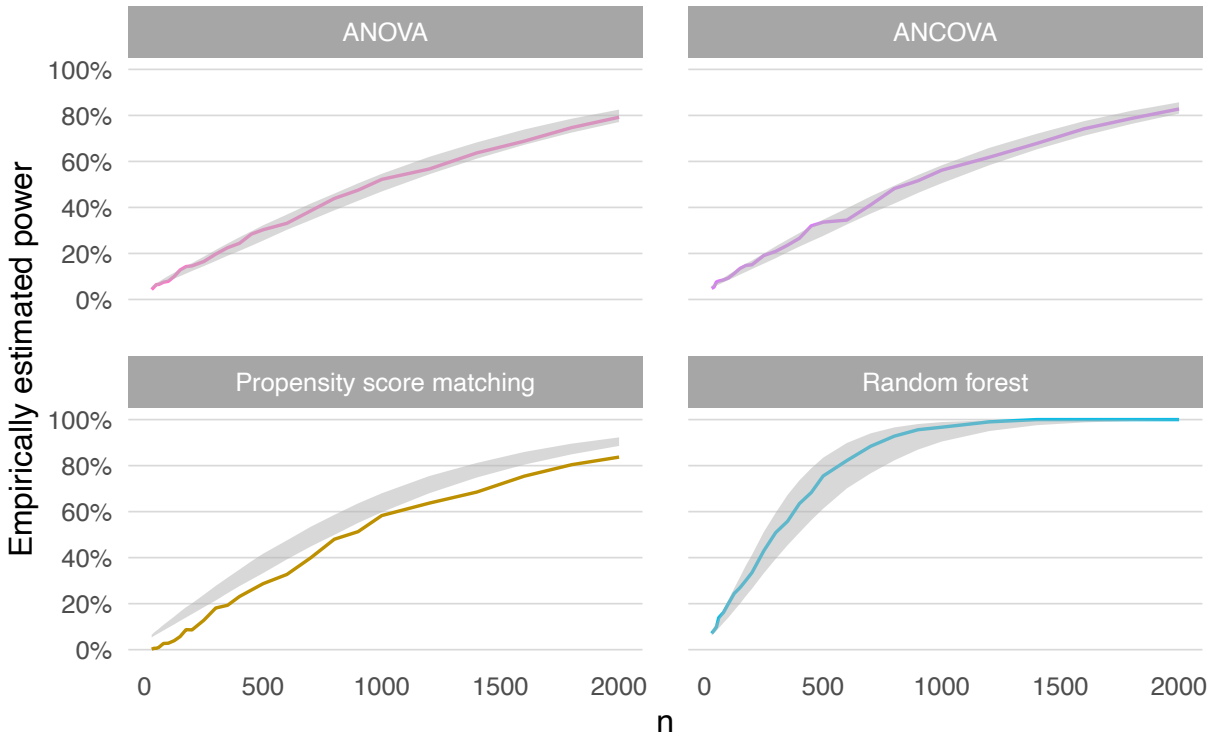


Figure D.2: Empirically estimated power together with the mean and the 2.5% and 97.5% quantiles of the prospective Guenter-Schouten power approximation in the case of a homogeneous treatment effect.

E Novo Nordisk A/S Clinical Trial Data

This appendix consists of information regarding the trial data presented in section 7.1 and 7.2.2.

E.1 Trial NN1218-3853

Table E.1 contains the inclusion and exclusion criteria for trial NN1218-3853.

E.2 Trial NN1218-4049

Table E.2 contains the inclusion and exclusion criteria for trial NN1218-4049.

E.3 Trial NN1250-3998

Table E.3 contains the inclusion and exclusion criteria for trial NN1250-3998.

E.4 Variables and their Distributions

Table E.4 contains a description of every covariate contained in the provided data sets along with the number of missing values in current and historical data.

Figures E.1 and E.2 show the empirical data distributions of all selected variables.

Inclusion criteria	Exclusion criteria
Informed consent obtained before any trial-related activities	Any use of bolus insulin, except short term use due to intermittent illness
Age $\geqslant 18$	Use of GLP-1 agonists and/or TZDs within the last 3 months prior to screening
Type 2 diabetes diagnosed clinically $\geqslant 6$ months at time of screening	Anticipated change in concomitant medication known to interfere significantly with glucose metabolism after randomisation
Treated with basal insulin for at least 6 months prior to screening	Cardiovascular disease within the last 6 months prior to screening
Current once daily treatment with insulin NPH, insulin detemir or glargine for at least 3 months prior to the screening visit	Systolic blood pressure $\geqslant 180 mmHg$ and/or diastolic blood pressure $\geqslant 100 mmHg$ after 5 minutes rest in a sitting position using a mean of 3 measurements
Current treatment with metformin with or without combination with other OADs at least 3 months prior to screening	Impaired liver function, defined as ALAT $\geqslant 2.5$ times upper limit of normal range
HbA1c between $7.0 - 9.5\%$ in the group receiving only metformin and between $7.0 - 9.0\%$ in the group receiving metformin in combination with other OADs	Imparied renal function defined as serum creatine $> 135 \mu mol/L$ for males and $> 110 \mu mol/L$ for females, or estimated creatinine clearence below $60 mL/min$
BMI $\leqslant 40 kg/m^2$	Recurrent severe hypoglycaemia or hypoglycaemic unawareness judged by the Investigator or hospitalisation for diabetic ketoacidoses during the previous 6 months prior to screening
Ability and willingness to adhere to the protocol	Proliferative retinopathy or maculopathy requiring treatment judged by investigator
Ability and willingness to eat at least 3 meals (breakfast, lunch, dinner) every day during the trial	Female of childbearing potential who are pregnant, breastfeeding or intend to become pregnant or are not using adequate contraceptive methods
Not currently using real time CGM system and/or consent to not use real time CGM system during trial other than the blinded one handed out in the trial if selected to the CGM subgroup	Any clinically significant disease or disorder, except for conditions associated with type 2 diabetes, which in the Investigator's opinion might jeopardise subject's safety or compliance with the protocol
	Any condition that the Investigator judges would interfere with evaluation of the results
	Mental incapacity, psychiatric disorder, unwillingness or language barriers
	Previous participation in this trial
	Known or suspected hypersensitivity to any of the trial products
	Donation of blood within 1 month prior to screening
	Known or suspected abuse of alcohol, narcotics or illicit drugs

Table E.1: Inclusion and exclusion criteria of trial NN1218-3853

Inclusion criteria	Exclusion criteria
Informed consent obtained before any trial-related activities	Any use of bolus insulin, except short term use due to intermittent illness
Age $\geqslant 18$	Use of GLP-1 agonists and/or TZDs within the last 3 months prior to screening
Type 2 diabetes diagnosed clinically $\geqslant 6$ months at time of screening	Anticipated change in concomitant medication known to interfere significantly with glucose metabolism after randomisation
Treated with once daily insulin detemir, insulin glargine or NPH for at least 3 months prior to screening	Cardiovascular disease within the last 6 months prior to screening
Current treatment with metformin (at least 1000 mg) with or without combination with other OADs at least 3 months prior to screening	Systolic blood pressure $\geqslant 180 mmHg$ and/or diastolic blood pressure $\geqslant 100 mmHg$ after 5 minutes rest in a sitting position using a mean of 3 measurements
HbA1c between $7.0 - 9.5\%$ in the group receiving only metformin and between $7.0 - 9.0\%$ in the group receiving metformin in combination with other OADs	Impaired liver function, defined as ALAT/SGPT $\geqslant 2.5$ times upper limit of normal range
BMI $\leqslant 40 kg/m^2$	Imparied renal function defined as serum creatine $> 135 \mu mol/L$ for males and $> 110 \mu mol/L$ for females, or estimated creatinine clearence below $60 mL/min$
Ability and willingness to adhere to the protocol including performance of SMPG	Recurrent severe hypoglycaemia or hypoglycaemic unawareness judged by the Investigator or hospitalisation for diabetic ketoacidoses during the previous 6 months prior to screening
Ability and willingness to eat at least 3 meals (breakfast, lunch, dinner) every day during the trial	Proliferative retinopathy or maculopathy requiring treatment judged by investigator
	Female of childbearing potential who are pregnant, breastfeeding or intend to become pregnant or are not using adequate contraceptive methods
	Any clinically significant disease or disorder, except for conditions associated with type 2 diabetes, which in the Investigator's opinion might jeopardise subject's safety or compliance with the protocol
	Any condition that the Investigator judges would interfere with evaluation of the results
	Mental incapacity, psychiatric disorder, unwillingness or language barriers
	Previous participation in this trial
	Known or suspected hypersensitivity to any of the trial products
	Donation of blood within 1 month prior to screening
	Known or suspected abuse of alcohol, narcotics or illicit drugs

Table E.2: Inclusion and exclusion criteria of trial NN1218-4049

Inclusion criteria	Exclusion criteria
Informed consent obtained before any trial-related activities	Known or suspected hypersensitivity to any of the trial products or related products
Age $\geqslant 18$	Previous participation in this trial
Type 2 diabetes diagnosed clinically $\geqslant 26$ weeks at time of screening	Female of childbearing potential who are pregnant, breast-feeding or intend to become pregnant or are not using adequate contraceptive methods
Treated with any basal insulin with or without OADs (metformin, DPP-4 inhibitor, alpha-glucodiase inhibitor, thiazolidinediones and SGLT2-inhibitor) for at least 26 weeks prior to screening	Treatment with a bolus insulin within the last 26 weeks prior to screening
HbA1c $\leqslant 9.5\%$	Use of any other anti-diabetic agents than those stated in the inclusion criteria within the last 26 weeks prior to screening
BMI $\leqslant 45 kg/m^2$	Receipt of any investigational medicinal product within 4 weeks prior to screening
Ability and willingness to adhere to the protocol including performance of SMPG	Any chronic disorder or severe disease, except for conditions associated with type 2 diabetes, which in the Investigator's opinion might jeopardise subject's safety or compliance with the protocol
Subjects fulfilling at least one of the below criteria: • Experienced at least one severe hypoglycaemic episode within the last year • Moderate chronic renal failure, defined as glomerular filtration rate $30-59 mL/min/1.73m^2$ per CKD-Epi • Hypoglycaemic symptom unawareness • Exposed to insulin for more than 5 years • Hypoglycaemic episode within the 12 weeks prior to screening	Uncontrolled or untreated severe hypertension defined as systolic blood pressure $\geqslant 180 mmHg$ and/or diastolic blood pressure $\geqslant 100 mmHg$ Impaired liver function, defined as ALAT or ASAT $\geqslant 2.5$ times upper limit of normal range Imparied liver function defined as glomerular filtration rate $< 30 mL/min/1.73m^2$ per CKD-Epi Proliferative retinopathy or maculopathy requiring acute treatment judged by investigator Stroke; decompensated heart failure New York Heart Association class III or IV; myocardial infarction; unstable angina pectoris; or coronary arterial bypass graft or angioplasty; all within the last 26 weeks prior to screening

Table E.3: Inclusion and exclusion criteria of trial NN1250-3998

Variable name	Variable description	Number of NAs	
		Current data	Historical data
Trial variables			
USUBJID	Subject id	0	0
STUDYID	Clinical trial id	0	0
SITEID	Medical site id	0	0
FASFL	Full analysis set population flag	0	0
Treatment allocation variables			
BOLUSP	Type of bolus insulin	0	0
BASALP	Type of basal insulin	0	0
METFORMIN	Did patient receive metformin?	0	0
OTHER OAD	Did patient receive some OAD other than metformin?	0	0
Demographics			
COUNTRY	Country	0	0
AGE	Age in years	0	0
SEX	Sex	0	0
RACE	Race	0	0
ETHNIC	Ethnicity (hispanic/latino or not)	0	0
General health measurements			
BMIBL	Body mass index (kg/m^2)	0	0
DIABTYP	Diabetes type (1 or 2)	0	0
DIABDUR	Number of years since diagnosis of diabetes	0	1
SMOKER	Frequency of smoking	0	1
SYSTOLIC	Systolic blood pressure (mmHg)	0	0
DIASTOLIC	Diastolic blood pressure (mmHg)	0	0
PULSE	Pulse (beats/min)	0	0
PEC	Number of physical examination abnormalities	0	2
CMC	Number of comorbidities	0	0

(continued on next page)

(continued from previous page)

Blood measurements

AG	1.5 Anhydroglucitol (μg/mL)	2	721
ALANINE	Alanine Aminotransferase Serum (U/L)	0	1
ALBUMIN	Albumin Serum (g/dL)	0	2
ALKALINE	Alkaline Phosphatase Serum (U/L)	0	1
ASPARTATE	Aspartate Aminotransferase Serum (U/L)	0	1
CREATININE	Creatinine Serum (μ-mol/L)	0	1
CrCl	Calculated Creatinine Clearance serum (mL/min)	0	1
FPG	Fasting plasma glucose (mmol/L)	0	15
HAEMATOCRIT	Haematocrit Blood (%)	0	6
HAEMOGLOBIN	Haemoglobin Blood (mmol/L)	0	3
HDL	HDL Cholesterol Serum (mmol/L)	0	721
CRP	High sensitive C-reactive protein (mg/L)	153	812
LDL	LDL Cholesterol serum (mmol/L)	0	721
ERYTHROCYTES	Erythrocytes (10^{12}/L)	0	3
LEUKOCYTES	Leukocytes (10^9/L)	0	3
LYMPHOCYTES	Lymphocytes (%)	0	721
MONOCYTES	Monocytes (%)	0	721
THROMBOCYTES	Thrombocytes (10^9/L)	0	15
NEUTROPHILS	Neutrophils (%)	0	721
BASOPHILS	Basophils (%)	0	721
EOSINOPHILS	Eosinophils (%)	0	721
NATRIURETIC	N-Terminal ProB-type Natriuretic Peptide (pmol/L)	153	805
POTASSIUM	Potassium Serum (mmol/L)	0	1
SMPG	Self Measured Plasma Glucose level (mmol/L)	0	720
SODIUM	Sodium Serum (mmol/L)	0	1
BILIRUBIN	Total Bilirubin Serum (μmol/L)	0	2
CHOLESTEROL	Total Cholesterol Serum (mmol/L)	0	721
PROTEIN	Total Protein serum (g/dL)	0	720
TRIGLYCERIDES	Triglycerides Serum (mmol/L)	0	721
HbA1cBL	HbA1c (%) at baseline	0	0
HbA1cET	HbA1c (%) at end of treatment	0	0
HbA1cCH	Change in HbA1c from baseline to end of treatment	0	0

Table E.4: Variables considered for the historical and current data sets. All variables except HbA1cET and HbA1cCH are measured at or before baseline. The number of current trial patients is $n = 153$, and the number of historical data patients is $n' = 1492$. Rows marked with grey are covariates removed due to many missing values.

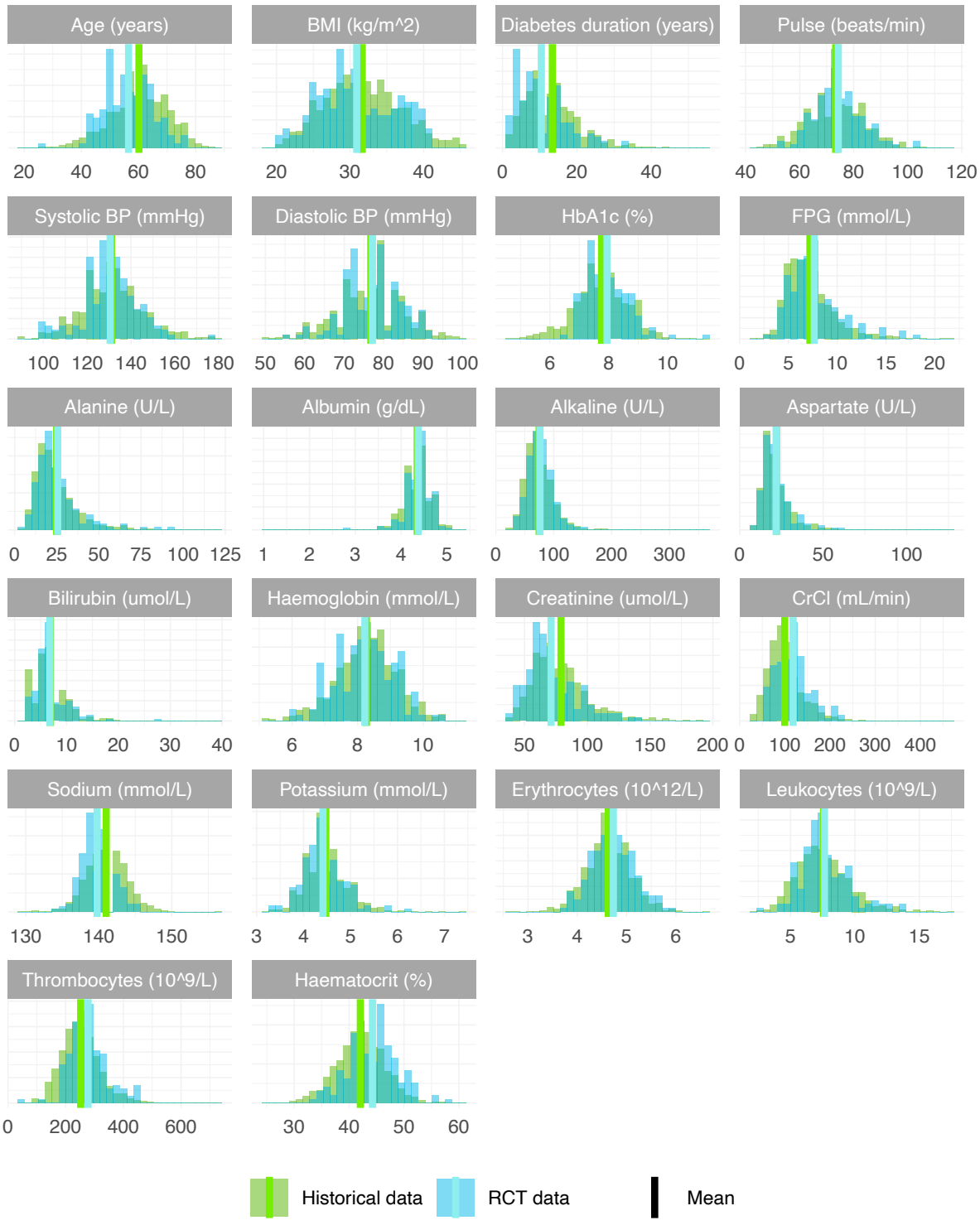


Figure E.1: Empirical distributions of continuous covariates in current and historical data. All measured at or before baseline. All measurements from and including HbA1c are obtained from blood samples. BP: Blood pressure. FPG: Fasting plasma glucose. CrCl: Creatinine clearance.

Figure E.2: Empirical distributions of discrete covariates in current and historical data. All measured at or before baseline. B/AA: Black or African American; AI/AN: American Indian or Alaska native; NW/OPI: Native Hawaiian or other Pacific Islander. ARG: Argentina; CAN: Canada; GBR: Great Britain; HRV: Croatia; IND: India; ISR: Israel; MEX: Mexico; ROU: Romania; RUS: Russia; SRB: Serbia; SVK: Slovakia; SVN: Slovenia; USA: United States of America.