Hypothetical Estimands in Randomised Controlled Trials: Unifying Causal Inference and Semiparametric Theory

Louise Østerby Jespersen
Silje Post Strøm

Master's thesis, Group MS10-01, Mathematical Statistics

Department of Mathematical Sciences
Aalborg University
Thomas Manns Vej 23
9220 Aalborg Øst
Denmark
http://math.aau.dk

AALBORG UNIVERSITY

MASTER'S THESIS

Title

Hypothetical Estimands in Randomised Controlled Trials: Unifying Causal Inference and Semiparametric Theory

Theme

Causal inference for time varying treatment

Project Period

September 2024 - May 2025

Project Group

Group MS10-01

Authors

Louise Østerby Jespersen
Silje Post Strøm

Supervisor

Rasmus Plenge Waagepetersen

Industrial co-supervisors

Claus Dethlefsen
Henrik F. Thomsen

Page Numbers

95

Date of Completion

May 24, 2025

Abstract

When targeting the hypothetical estimand in a randomised controlled trial, accounting for intercurrent events in the analysis presents significant challenges as intercurrent events have a confounding effect. This project presents the causal inference workflow in the context of randomised clinical trials. In addition, the project presents the theory of semiparametric models in order to present the targeted learning framework. When determining the efficacy of treatments in terms of the hypothetical estimand, common practice is to use a Mixed Model for Repeated Measures (MMRM). This project proposes the use of Longitudinal Targeted Maximum Likelihood Estimation (LTMLE) for estimating the hypothetical estimand. Through simulations and empirical analysis, we assess how these methodologies manage the impact of varying amounts of intercurrent events on treatment outcomes. Our findings suggest that while MMRM provides an easily interpretable solution, LTMLE offers a more robust solution by more accurately reflecting the causal relationships in the presence of rescue medication and treatment discontinuation.

Preface

This master's thesis has been completed as part of our studies at the master in Mathematics at Aalborg University during the 3rd and 4th semesters. Throughout this period, we undertook a three-month stay at the University of California, Berkeley, where we were kindly invited by Professor Maya Petersen. During this stay we attended a class in Targeted Learning taught by Professor Mark van der Laan, whom we would like to express our sincere gratitude to, for his invaluable assistance and teaching during our time abroad. Additionally, we would like to thank Christophe Biscio, Rasmus Plenge Waagepetersen, Claus Dethlefsen, Christina da Silva, and Jessica Angel for their support in making this trip possible and successful. The trip was made possible with financial support from the William Demant Foundation, the Augustinus Foundation, and the Danish Data Science Academy.

Our project is rooted in a practical problem, and we are particularly grateful to Novo Nordisk A/S for providing us with data that has significantly enriched our analysis and findings. Working on this project for a year has been an educational process. We have found ourselves navigating in deep waters, as we have explored an entirely new subject with significant potential for the future. It has been intriguing to be part of a process where we were allowed to pursue what interests us, while focusing on learning as much as possible along the way. Simultaneously, we have improved our ability to set certain limitations for ourselves and accepted that we cannot explore every aspect of our research topic.

Last but not least, we extend our heartfelt thanks to our supervisors, Rasmus Plenge Waagepetersen from the University, Claus Dethlefsen, and Henrik Thomsen from Novo Nordisk A/S, for their guidance and support throughout this journey. As we complete our master's program, we remain dedicated to lifelong learning and look forward to the enriching experiences that will shape our continued educational endeavours.

In our thesis, tables, figures, and examples are numbered according to their respective chapters, while definitions, theorems, and equations are numbered according to their sections.

Aalborg University, May 24, 2025

Contents

List of abbreviations and notation

Table 1: Abbreviations that are frequently used throughout the project are listed here.

Abbreviation	Expansion
RCT	Randomised controlled trial
ATE	Average treatment effect
ICE	Intercurrent Event
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
EMA	European Medicines Agency
FDA	The U.S. Food and Drug Administration
PIONEER	Peptide InnOvatioN for Early diabEtes tReatment
HbA_{1c}	Glycosylated haemoglobin A1c
T2D	Type 2-diabetes
FPG	Fasting plasma glucose
DGP	Data generating process
SCM	Structural causal model
IF	Influence function
EIF	Efficient influence function
CLT	Central limit theorem
TMLE	Targeted Maximum Likelihood Estimation/Estimator
LTMLE	Longitudinal Targeted Maximum Likelihood Estimation/Estimator
BL	Baseline
SAP	Statistical Analysis Plan
MMRM	Mixed Model for Repeated Measures
RMSE	Root mean squared error.

Table 2: Notation that is frequently used throughout the project is listed here.

Notation	Description
$\mathbb{N}$	The natural numbers, excluding zero.
$\mathbb{R}$	The real numbers.

Continued on next page

Table 2: Notation that is frequently used throughout the project is listed here. (Continued)

Notation	Description
I_n	The $n \times n$-dimensional identity matrix.
$\mathbb{0}_j$, $\mathbb{1}_j$	The j-dimensional vectors of zeros and ones, respectively.
W (W_0)	Vector of baseline covariates.
W_k	Measurement of the outcome variable at visit k, for $k = 1, 2, \ldots$.
$\mathbb{W}_k$	All potential outcomes of W_k.
A (A_0)	Randomised treatment.
A_k	Indicator of treatment discontinuation at visit k, for $k = 1, 2, \ldots$.
Z	Indicator of rescue medication initiation.
Z_k	Indicator of rescue medication initiation at visit k, for $k = 0, 1, \ldots$.
$V_{j:k}$	The vector of variables $(V_j, \ldots, V_k)$ for $j \leq k$.
Y	Outcome.
$Y(a)$	Potential outcome where $A = a$.
$Y(a, z)$	Potential outcome where $A = a$ and $Z = z$.
$\mathbb{Y}$	All potential outcomes of Y.
P_0	True probability distribution of the observed data, that is the true DGP.
P^*	True probability distribution for the data structure involving potential outcomes.
P	Possible DGP.
p_0	Density of true DGP P_0.
p	Density of P.
Pa(V)	Parents of the variable V.
$\mathcal{M}$	Statistical model, which is a collection of possible DGPs for the observed data.
$\mathcal{M}^*$	Causal model, which is a collection of possible DGPs for the data structure involving potential outcomes.
$\Psi(\cdot)$	Statistical estimand.
$\Psi^*(\cdot)$	Causal estimand.
$\Psi(P_0), \Psi^*(P^*)$	Target parameters.
ϕ_{P_0}	Influence function of an estimand.
$\bar{Q}_{W_k}$	Mean of the conditional distribution of W_k given Pa(W_k).
$\bar{Q}_Y$	Mean of the conditional distribution of Y given Pa(Y).
$G(W)$	Mean of the conditional distribution of A given W.
g_{V_k}	The density of the conditional distribution of V_k given Pa(V_k)
$\mathcal{N}_n(\mu, \Sigma)$	The n-dimensional normal distribution with mean μ and variance matrix Σ.
t_ν	The univariate t-distribution with ν degrees of freedom.
$\xrightarrow{d}$	Convergence in distribution.
$\xrightarrow{P}$	Convergence in probability.

1 Introduction

A phase three *randomised controlled trial* (RCT) is generally conducted with the purpose of demonstrating the efficacy and safety of an experimental treatment. It is often conducted because of some questioning by experts in clinical practice. RCTs are considered the gold standard for drug approval as randomisation implies some very favourable properties in causal inference. Randomisation is crucial in these trials as it guarantees that the distribution of both observed and unobserved baseline covariates and hence confounding factors is similar across treatment groups. However, participants in the study may experience worsening of symptoms, insufficient therapeutic effects or unforeseen complications. In response to this, they might start non-randomised disease targeting medication in addition to the randomised treatment [1]. This non-experimental treatment will be referred to as *rescue medication* in this project. Some participants may even discontinue the randomised treatment in response to complications. Receiving rescue medication and discontinuing treatment are examples of an *intercurrent event* (ICE), which is a term used for events that occur after randomisation that may affect the assessment and interpretation of the outcome [2]. ICEs can be related to the treatment or the disease, but sometimes it is completely unrelated to either.

ICEs occur in almost all trials and sponsors are usually encouraged to record them as part of the trial conduct. Naturally, this data is used for assessing safety of a treatment, but it is also important for showing efficacy as intercurrent events often affect the outcome. In [3], ICEs are an important part of answering the clinical question of interest, which will be discussed in greater detail in Section 1.2. As a motivation for why this discussion on ICEs is so important, we will exemplify the setting by considering rescue medication. We may expect that rescue medication will affect the outcome in some way as it is additional medication that is meant to treat the same disease as the experimental treatment. Rescue medication is only given to participants in need of it making it a non-randomised treatment. As a consequence, this may introduce some bias as the need for rescue medication might be uneven across treatment arms. This is of course concerning as most trials are set up as randomised trials to avoid bias. Hence, the use of rescue medication complicates the interpretation and analysis of the trial results if one wishes to determine the treatment effect in the hypothetical scenario where rescue medication was not available. This is of interest, since the effect of the experimental treatment in the case where it is not influenced by other medications is important for documentation of the experimental drug's efficacy.

It may seem like an easy fix to design the study such that participants do not end up taking rescue medication or discontinue the randomised treatment by simply forbidding it. However, this approach would be highly unethical as the participant's safety should be of highest priority when conducting a trial. In addition, there are ICEs like for example death that would not be possible to hinder no matter what. Hence we need methods which are able to account for the occurrence of ICEs. Common practice for determining the treatment effect in the scenario where ICEs are absent, has been to discard all samples collected after any ICE, and therefore handle them as missing. However, it is of interest to explore other methods that utilise all the collected information and

models the effects of ICEs, instead of disregarding observations that are affected by ICEs. This approach is especially of interest in studies involving chronic and rare diseases as rescue medication intake is a lot more common making very little data available if we disregard all the affected data.

1.1 Causal inference workflow

Before conducting a trial it is extremely important to make some statistical considerations and decisions with the objective in mind. The aim of incorporating this into the planning of a trial is to ensure that the trial data are reasonable to base an answer upon. This can be ensured by following the workflow presented in this section, going from the clinical world into the statistical world through the world of causality.

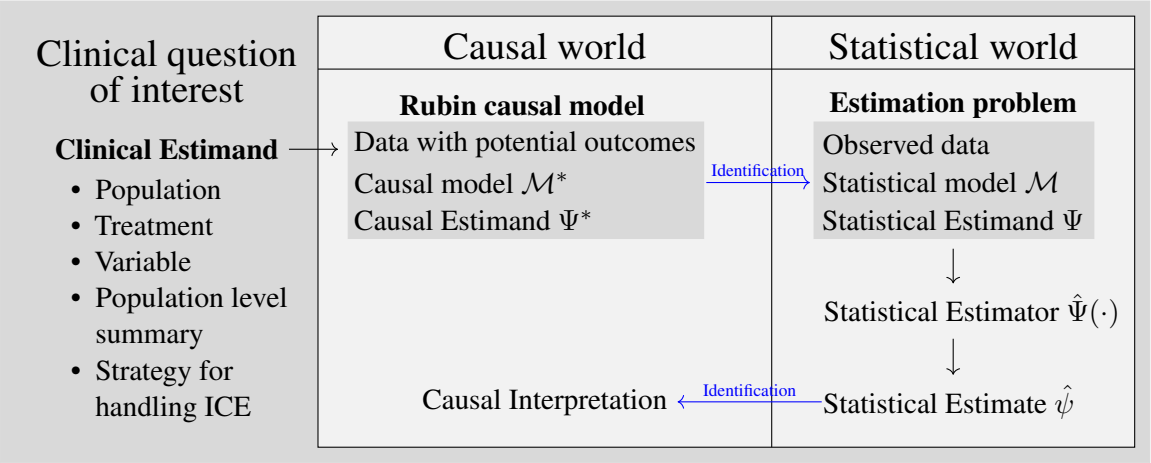


Figure 1.1: *Causal inference workflow that gives an overview of the different types of estimands and how they relate to each other.*

Figure 1.1 displays the ideal workflow for answering clinical research questions based on clinical trials. We start on the left-hand side, where we have the clinical question of interest which are specified to align the goals of conducting the study. This clinical question should now be translated into a clinical estimand, which will be introduced in Section 1.2, explicitly outlining the five attributes to reduce uncertainties in the answer to the clinical question. Since clinical practice and proper statistics are completely different, this clinical estimand serves as a tool to connect the clinician's knowledge and the expertise of the statistician. When this has been decided, we can move on to considering the Rubin causal model, which will be introduced in Subsection 2.2.1. Here, the five attributes of the clinical estimand are formulated mathematically in terms of the causal data structure, causal model and causal estimand. Intuitively, one can think of defining the causal estimand as answering the question "In an ideal world where we can observe the outcomes in either treatment scenario, what quantity would we like to find?".

In order to move on to the statistical world on the right-hand side, we rely on an identification result under some identifiability assumptions which will be described in Subsection 2.2.4 and is depicted with a blue arrow. The observed data, the statistical model and the statistical estimand make up the *estimation problem*, which can be solved using statistical methods.This leads to formulating a statistical estimator that produces a statistical estimate, which will be the topic of Chapter 3. After this process, we again rely on the identifiability result, hence the blue arrow, to make a causal interpretation of the estimate. The causal interpretation makes us able to say something about the ideal world where we can observe outcomes in either treatment scenario, based only on observed

data. At this point our job as statisticians is done if we can formulate this causal interpretation in a way where stakeholders are able to understand it and use it to make important decisions.

1.2 Handling of intercurrent events

As mentioned earlier, ICEs affect the observed outcome and hence the interpretation of the results based on the outcome. Hence, if ICEs can not be avoided in clinical trials, it is important to handle them properly. In this section, we will stress the importance of accounting for ICEs by using an example and afterwards discuss some of the recommendations made by the regulatory authorities about this matter.

Imagine that we conduct a trial, where the effect is measured in an outcome variable measured repeatedly through time, and a lower value corresponds to an improvement. On the other hand, a too high value of the outcome variable makes a participant more prone to receiving rescue medication, due to safety concerns. Figure 1.2, depicts the mean progression of some biomarker of the participants in either treatment group throughout time together with a grey area indicating when a participant will receive rescue medication. The two groups are equal in the mean value of their biomarker at the beginning of the trial, but the treatment group improves more over time, compared to the placebo group. This implies that later on in the trial placebo participants are much more prone to receiving rescue medication, since they are more likely to enter the grey area. As a result, the biomarker is lowered in the placebo group too. This is explained by the fact that rescue medication is often defined as other disease-targeting medication. At the end of the trial, we see a difference between placebo and the treatment group, depicted by the blue arrow. But, in addition we have depicted, by a dashed red line, how the placebo group would have behaved, had rescue medication not been available. Following this dashed line, we see that the actual difference, depicted by a yellow arrow, had rescue medication not been available, is bigger than when rescue medication is allowed. Hence, if one does not account for the intake of rescue medication in the statistical analyses, the treatment effect could be underestimated. This is especially a concern in cases where the intake of rescue medication is not balanced between the treatment groups.

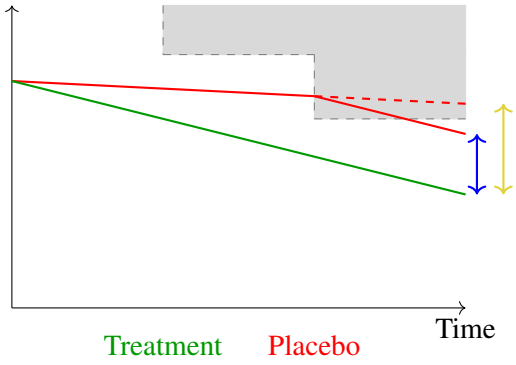


Figure 1.2: *This figure illustrates the mean trajectory of some biomarker showing disease progression in treatment group (green line) and placebo group (red line) over time. The blue vertical arrow represents the observed changes in the biomarker, highlighting the difference in treatment effectiveness over time compared to placebo. The shaded area illustrates the threshold of the biomarker for receiving rescue medication and the dashed red line indicates the progression of the placebo arm had rescue medication not been available. In this case, the yellow vertical arrow represents the the difference in treatment effectiveness over time compared to placebo.*

One very important aspect in the handling of data obtained after occurrence of an ICE in a clinical

trial is what the regulatory authorities suggest. This is important since their approval is crucial in the process of getting a new treatment on the market.

The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) in cooperation with The European Medicines Agency (EMA) published an addendum called *Estimands and sensitivity analysis in clinical trials* [3] to the overall statistical guideline *ICH E9 Statistical Principles for Clinical Trials* [4]. In general, this addendum states that it is important to prespecify strategies for handling ICEs. This is done through specifying a *clinical estimand*, corresponding to the clinical question of interest, which has five attributes:

- Target population
- Treatment condition of interest
- Outcome variable of interest
- Population-level summary providing a basis for treatment group comparison
- Strategies for handling ICEs

Each of these attributes influence the interpretation of the results obtained from the analysis. The purpose of specifying each of these five attributes is to eliminate uncertainties in the clinical question. For examples of how one can specify these attributes see [5].

In addition to this, different strategies for handling data collection after occurrence of an ICE is outlined in the addendum. In the following, we will highlight the strategies that will be considered in this project:

- *Treatment Policy strategy*: Using this strategy we will disregard the occurrence of ICEs and use the values of the affected variables in the analysis as if they had not been affected.
- *Hypothetical strategy*: Considering a hypothetical scenario for the ICE. This could be the scenario that the participant received the assigned treatment and not experienced any ICE.

The treatment policy estimand will report the estimated treatment effect on a population level. That is, the effect that we would observe if we released this drug in to the population. The treatment policy estimand does not take any events between randomisation and observing the outcome into account. Often the authorities are interested in this effect, as it gives an idea of what happens to the population if we just release this medication and people did whatever they want with it. In contrast, the hypothetical estimand will report the estimated treatment effect on an individual level and it "*is relevant for physicians who base clinical patient-specific decisions on the anticipated achievable treatment effect*" [6]. This effect applies to individuals that do not experience any of the ICEs that are handled using the hypothetical strategy. When a patient is diagnosed with a disease the patient does not care about the population level effect of the prescribed medication which is contaminated by non-adherence, protocol deviation and additional medication. Instead the hypothetical estimand reports the effect of interest for the patient. It is important to notice that the guidelines solely emphasise the usage of estimands, not which strategy to use or that one strategy is better than the other. Often, the two estimands, the treatment policy estimand and the hypothetical estimand, augmented with the two specified strategies for handling ICE, respectively, will complement each other in analyses. However, one of the estimands has to be the primary one, at least in medicine, since they have to report one effect on the label.

1.3 Brief introduction to PIONEER 1

The PIONEER phase 3a clinical development programme for oral semaglutide is a global development programme with type 2 diabetes patients which completed in 2018. We have been provided access to data from the PIONEER 1 trial by Novo Nordisk A/S. In this trial, ICEs like discontinuation of trial product and all additional medications were registered. Most of the time endpoint measurements of HbA_{1c} were collected even after these ICEs occurred, making the resulting data appropriate for looking into the hypothetical strategy for handling this ICE.

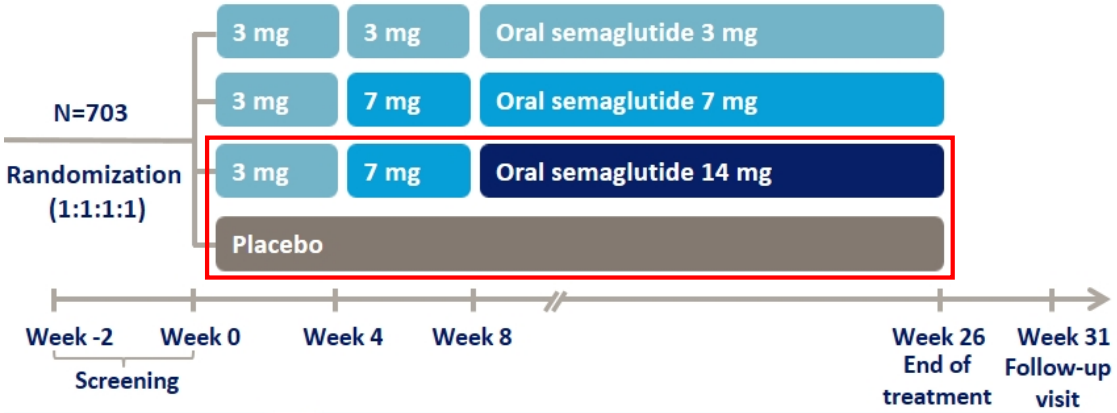


Figure 1.3: *Trial design for PIONEER 1 [6]. The red box highlights the arms that we will consider in this project.*

The design of the trial is illustrated in Figure 1.3, showing the exact number of randomised participants, 703, as opposed to the planned number of participants, 704. The duration of the trial was approximately 33 weeks, from screening visit, 2 weeks before randomisation, to the follow-up visit, 5 weeks after end of treatment. At the screening visit participants were assessed for their eligibility according to inclusion and exclusion criteria. One of the inclusion criteria was that the participants need to be treated with diet and exercise only 30 days prior to screening. To see full inclusion and exclusion criteria for PIONEER 1, see Table B.1. The trial is a four-armed study with a randomisation ratio of 1:1:1:1. Participants allocated to each of the three active treatment arms, displayed as the top three rows of boxes in Figure 1.3, start on the lowest dose, 3 mg, and then gradually increase the dose up to their respective target doses. This is illustrated by colour coded boxes sorted in columns to make it clear that the first increase, assuming that the target dose is not reached, is at week 4 and the next at week 8. For simplicity, we have chosen to focus on comparing placebo and the arm with 14 mg as target dose, corresponding to the two bottom rows of boxes in Figure 1.3, marked by a red rectangle. Hence, when using the trial data in this project, we will consider it as a two armed study with randomisation ratio 1:1.

The primary endpoint when seeking for efficacy is the change in glycosylated haemoglobin A1c (HbA_{1c}) from baseline to end of treatment at week 26. The variable HbA_{1c} reflects the average blood sugar level over the past two to three months, often expressed in percentage [7]. The range in which the blood sugar level is considered normal for adults is $HbA_{1c} < 5.7\%$. When HbA_{1c} is between 5.7% and 6.4% it is considered so-called pre-diabetes, and then diabetes when it is above 6.5%. In Table B.1, which shows the inclusion criteria, it should be noted that patients must have an HbA_{1c} above 7.0% in combination with at least 3 months of diagnosed diabetes, to be considered eligible to participate in the trial. This biomarker is a good indicator of overall glycaemic control. In patients with diabetes, it is desirable to lower HbA_{1c} since this will indicate well-regulated long-term diabetes.

As an other measure of glycaemic control, the fasting plasma glucose (FPG) is also measured at each visit. FPG is the glucose level in the blood after at least 8 hours of fasting. This gives an insight into the participants glycaemic control at the very moment that the measurement is taken. Naturally, FPG measurement have a lot more variability than HbA_{1c} measurements making it a poor endpoint for determining efficacy for a treatment. However, it is important to measure to monitor the safety of the participants and use as a guidance for whether participants are in need of any additional anti-diabetic medication. This will be discussed in greater detail in Section 5.1.

Semaglutide is a glucagon-like peptide-1 (GLP-1) analogue treatment for patients with type 2 diabetes. It lowers blood glucose levels by stimulating the secretion of insulin and reducing the secretion of glucagon [8]. This treatment is crucial for patients diagnosed with diabetes, since a too high level of blood sugar, called hyperglycaemia, in an extended period of time comes with a lot of potential side effect. For example it can lead to health problems that affect the eyes, kidneys, nerves and heart, and it can cause serious health problems that require emergency care [9]. It can potentially lead to ketoacidosis, which is life threatening. On the other hand, one of the side effects in the treatment of diabetes is too low level of blood glucose, potentially causing unconsciousness. Hence, maintaining a stable level of blood glucose is essential for a healthy living.

Based on this case we can illustrate the first step in the workflow, Figure 1.1. We start from the clinical question of interest, which is

What is the treatment effect of oral semaglutide 14 mg compared to placebo on patients with T2D in the hypothetical scenario where rescue medication was not available and no one discontinues trial product? [6]

Now we can more formally define this in terms of attributes to the clinical estimand:

- **Population:** Participants aged ≥ 18 years with T2D for at least 30 days, who at trial entry were being treated with diet and exercise only and had HbA_{1c} levels between 7% and 9.5%, see Table B.1.
- **Treatment:** Oral semaglutide 14 mg vs. placebo.
- **Outcome variable:** Change in HbA_{1c} from baseline after 26 weeks.
- **Population level summary:** Mean difference in change in HbA_{1c} from baseline between treatment arms.
- **Strategy for handling ICE:** Hypothetical strategy for treatment discontinuation and rescue medication use.

We will return to this example later, and show the journey through the workflow, by illustrating a practical application of the steps. But since we are moving into the causal and statistical world, we need some theory before applying it.

In this project, we will mainly focus on discontinuation of trial product and initiation of rescue medication. Hence we will ignore the occurrence of any other ICE. However, the methodology discussed in this project aims to be general enough to apply to other ICEs. In Chapter 2, we will only focus on initiation of rescue medication to illustrate important concepts, which we will expand to a more complex case in Chapter 5 that takes both discontinuation of trial product and initiation of rescue medication into account.

2 Causal inference

This chapter will begin by introducing what we can obtain from experiments, data. Then it will move on to the causal world, where we explain causal models and causal estimands under Rubin's potential outcome framework. These will serve as the mathematical translation of the clinical estimand. The last part in Figure 1.1, going from the causal world to the statistical world is established by a section on identification. The concepts will be illustrated both by examples and graphics. It is mainly based on [10, ch. 1, 2, 3], [11, ch. 1, 2] and [12, ch. 1, 2].

2.1 Data

This section describes the importance of having some knowledge about the origin of the data and potential dependencies among variables, since "*data are meaningless without knowledge about the experiment that generated the data*" [11]. It takes inspiration from [13] in addition to the sources mentioned above.

Clinical questions often focus on specific properties of a population, like how they would respond to a given treatment. Ideally, we would have complete information on every single unit in this population and then be able to measure the exact property of interest. However, this is not feasible in practice, and hence we instead try to draw conclusions using a sample from the population of interest. Obtaining a valid sample involves many considerations that will impact the inference that one can draw from the sample. However, this is beyond the scope of this project as we have no say in the collection process of the data that we will consider in this project, which is introduced in Section 1.3.

Definition 2.1 (Data generating process)**.** A sample from the population of interest will be assumed to be n i.i.d. realisations of a stochastic vector that follows an unknown probability distribution which is referred to as the *data generating process* (DGP).

In statistics, we are interested in learning something about this DGP through the sample that has been collected. It is extremely important that we have some prior knowledge about the data and the experiment that it originates from to make it useful, which is described in the following.

Consider a collection of n i.i.d. realisations of a stochastic vector. *Specifying the data structure* in a study consists of a description of each variable in the vector, including its class and what it measures and the order in which the variables are observed with respect to each other, referred to as *time ordering*. This will be illustrated by the following example.

Example 2.1 Specification of the data structure
Consider the PIONEER 1 trial, presented in Section 1.3, which is an RCT where we collect data $o_i,\ i = 1, \ldots, n$, which are assumed to be n i.i.d. realisations of the stochastic vector $O = (W, A, Z, Y)$. We explicitly specify the data structure:

- W contains the baseline covariates sex (binary), region (categorical) and blood glucose level measured as HbA_{1c} (continuous).
- A is the randomised treatment allocation (binary), where $A = 0$ indicates that the participant is assigned to placebo and $A = 1$ indicates assignment to an anti-diabetic drug.
- Z is the binary indicator of whether a participant has initiated rescue medication throughout the study, where $Z = 1$ indicates initialisation of rescue medication at some point.
- Y is the blood glucose level measured as HbA_{1c} (continuous) after $T \in \mathbb{N}$ weeks, which serves as the outcome of interest.
- We observe the baseline covariates W prior to allocation to treatment A. Hereafter the indicator of rescue medication Z is recorded, and then the outcome Y is observed.

We will use the notation and data structure introduced in Example 2.1 throughout the entire project to exemplify different concepts.

Definition 2.2 (Confounder)**.** A variable which is a common cause of both the treatment and the outcome variable is called a *confounder* in the treatment-outcome relationship.

In an RCT the treatment allocation is random and hence independent of covariates, which implies that there will be no confounders in the treatment-outcome relationship. As data structures get more complicated, it is not always clear which variable is considered a treatment and which is considered an outcome. Hence, it may be important to specify the relationship that a variable is a confounder in. To illustrate the concept of a confounder the following example will be based on an observational study.

Example 2.2 Confounder
Consider an observational study aiming to determine the effect of smoking on lung cancer. In this case the treatment A could be a binary indicator of whether or not a person is smoking, and the outcome of interest Y is a binary indicator of whether or not the person has experienced lung cancer. A potential confounder in the A-Y relationship, which we will denote by W, is whether or not the participant's parents smoked. Passive smoking is believed to affect the risk of lung cancer, and having parents that smoke might influence on participants' own smoking habits.

It will later be evident that it is important to measure confounders, if there are any, and to incorporate them in the analysis.

2.2 Models and estimands

As explained, we will often have some prior knowledge of the way the data has been collected. This knowledge is very powerful in narrowing down the collection of potential DGPs and thereafter determining the quantity of interest, which are topics in this section. In addition to the sources mentioned earlier, this section is based on [14].

2.2.1 Rubin causal model

Causal questions are often related to a hypothetical world that we are unable to observe data from. For example, in a clinical study one might be interested in what would have happened if participants received a different treatment or no treatment at all, to explore the therapeutical effect.

To formalise this framework, we introduce the concept of *potential outcomes* as we will be working under the Rubin causal model in this project [15]. We start by introducing the Rubin causal model corresponding to the data structure described in Example 2.1 with inspiration from [12, sec. 1.2]. The potential outcome $Y(a, z)$ will denote the outcome whenever $A = a$ and $Z = z$. Potential outcomes answer the question of what would have happened to the participant's outcome in the different treatment scenarios. The trouble is that as a participant can not receive all treatment combinations simultaneously, only one of these potential outcomes can be observed for each participant in practice. This is known as the *fundamental problem of causal inference*.

Definition 2.3 (Causal model)**.** A *causal model* $\mathcal{M}^*$ is a collection of possible DGPs for a stochastic vector involving potential outcomes.

For the data structure $O = (W, A, Z, Y)$ introduced in Example 2.1, the causal model $\mathcal{M}^*$ will contain all possible DGPs for the stochastic vector $(W, A, Z, Y(0,0), Y(0,1), Y(1,0), Y(1,1))$. After defining the causal model $\mathcal{M}^*$, we need to determine the causal estimand which answers the clinical question of interest, in terms of potential outcomes.

Definition 2.4 (Causal estimand)**.** The *causal estimand* is a mapping

$$\Psi^* : \mathcal{M}^* \to \mathbb{R} \tag{2.2.1}$$

that associates each distribution in the causal model $\mathcal{M}^*$ with a number.

In words, the causal estimand is a mathematical formulation of the clinical estimand, incorporating each of the five attributes, see Section 1.2. That is, in the hypothetical world where we can observe all potential outcomes simultaneously, what quantity answers the clinical research question? The concept of a causal estimand can easily be extended to a higher dimensional mapping, which takes values in $\mathbb{R}^d$ for $d \in \mathbb{N}$, but it is out of scope for this project.

Example 2.3 (Continuation of Example 2.1) Rubin causal model
We are interested in determining the efficacy of the treatment A but without the influence of rescue medication Z, as explained in Section 1.3. Now, we can formulate the clinical estimand as a mathematical quantity in the causal world, where we assume that we can observe all potential outcomes for each participant and hence answer the question directly.

Rubin causal model

Data	Let $o_1^*, \ldots, o_n^*$ denote n i.i.d. observations from the stochastic vector $O^* = (W, A, Z, Y(0,0), Y(0,1), Y(1,0), Y(1,1))$. The covariates W and the potential outcomes are continuous measurements of HbA_{1c} bounded between 0 and 100%. W is measured at baseline and the potential outcomes are measured at the end of treatment. A is a binary indicator of randomised treatment, which is assigned right after measuring W, but is independent on W. Z is an indicator of whether or not a participant has taken rescue medication, which is measured after randomisation but before observing the final outcome.
Model	A causal model $\mathcal{M}^*$, which is a collection of distributions of O^* designed to comply with the natural constraints implied by data, eg. boundedness and independence.
Estimand	The causal estimand of interest is $$\Psi^*(P) = E_P[Y(1,0) - Y(0,0)]$$ for $P \in \mathcal{M}^*$, which corresponds to the clinical question of interest.

As a disclaimer, there are many different options for defining causal estimands other than the ATE. One might for example be interested in the covariance, density, population mean or the conditional average treatment effect, however these other possible causal estimands will not be discussed in this project. This section has now covered the middle part of Figure 1.1. Under the assumption of identification, the blue arrow in the figure, we are now able to move on to the statistical world, which is the topic in the next section.

2.2.2 The statistical estimation problem

A natural assumption is, that given the treatment A, the rescue medication indicator Z and the potential outcomes, the outcome Y that we actually observe is given by

$$Y = \sum_{a,z} I(A = a, Z = z) Y(a, z).$$

Hence the statistical part is simply the observed part, which is what we will cover in this section.

Definition 2.5 (Statistical model)**.** A *statistical model* $\mathcal{M}$ is a collection of possible DGPs for a observable stochastic vector.

From now on, we will assume that all probability distributions in $\mathcal{M}$ are dominated by a common measure μ, such that we can uniquely identify each possible probability measure $P \in \mathcal{M}$ by its density $p = \frac{dP}{d\mu}$. This is a rather technical assumption, which is necessary for ensuring that each probability distribution has a unique density, but we will not go into further details, but only consider models containing probability distributions that are dominated by a single measure.

Example 2.4 Specification of a statistical model
Consider a one-dimensional continuous stochastic variable X with an unknown distribution. If prior knowledge, on the experiment that generated observations from X suggests that the distribution may be normal, then the statistical model is

$$\mathcal{M} = \left\{ P \text{ with density } p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \;\middle|\; \text{for } \mu \in \mathbb{R}, \sigma \in \mathbb{R} \right\}. \tag{2.2.2}$$

This statistical model places an assumption on X that is quite restrictive as there are many other distributions that a continuous variable might follow. However, if we know that it is normally distributed, but don't know the parameters, this restriction makes sense as we now have a much smaller model space to search for the true DGP.

We need to make sure to avoid assumptions that are unrealistic for the data as this could possibly restrict the statistical model so much that the true DGP is no longer part of it.

Definition 2.6 (Statistical estimand)**.** The *statistical estimand* is a mapping

$$\Psi : \mathcal{M} \to \mathbb{R} \tag{2.2.3}$$

that associates each distribution in the statistical model $\mathcal{M}$ with a number.

Example 2.5 (Continuation of Example 2.3) Statistical estimand
We are interested in determining the efficacy of the treatment A but without the influence of rescue medication Z, as explained in Section 1.3. Now, we can formulate the Rubin causal model as a statistical estimation problem.

The statistical estimation problem

Data	Let $o_1, \ldots, o_n$ denote n i.i.d. observations from the stochastic vector $O = (W, A, Z, Y)$. W and Y are continuous measurements of HbA_{1c} bounded between 0 and 100%, measured at baseline and end of treatment respectively. A is a binary indicator of randomised treatment, independent on the baseline variable W. Z is an indicator of whether or not a participant has taken rescue medication. Y is equal to $Y(a, z)$ whenever $A = a$ and $Z = z$.
Model	A statistical model $\mathcal{M}$, which is a collection of distributions designed to comply with the natural constraints implied by data, eg. boundedness and independence.
Estimand	The statistical estimand of interest is $\Psi(P) = E_P[E_P[Y \mid A = 1, Z = 0, W] - E_P[Y \mid A = 0, Z = 0, W]]$. It will be clear later how this aligns with the causal estimand of interest and the clinical question of interest.

It is crucial, that there is a one-to-one correspondence between the statistical and causal world. We can guarantee this, under some non-testable assumptions, which we will return to and establish in Subsection 2.2.4.

2.2.3 Structural causal models and directed acyclic graphs

As a tool for illustrating the dependencies among variables, we will introduce the concept of a structural causal model.

Definition 2.7 (Structural causal model)**.** Consider a collection of time-ordered stochastic variables $X_1, \dots, X_K$ for $K \in \mathbb{N}$. Let $U = (U_1, \dots, U_K)$ be independent stochastic variables with joint distribution P_U. A *structural causal model* (SCM) consists of K structural assignments

$$X_k = f_k(S_k, U_k) \tag{2.2.4}$$

and the joint distribution P_U, where f_k is a function and $S_k \subseteq \{X_1, \dots, X_{k-1}\}$ for $k = 1, \dots, K$.

Let us consider an SCM for the stochastic vector $O = (W, A, Z, Y)$ introduced in Example 2.1. We can express the dependencies between the variables with respect to the time ordering in a system of structural equations

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Z &= f_Z(W, A, U_Z) \\ Y &= f_Y(W, A, Z, U_Y). \end{aligned} \tag{2.2.5}$$

Here $f = (f_W, f_A, f_Z, f_Y)$ represents the generating functions for the variables W, A, Z and Y, whereas $U = (U_W, U_A, U_Z, U_Y)$ are stochastic variables with joint distribution P_U. The variables W, A, Z and Y are stochastic through U_W, U_A, U_Z and U_Y, which in turn are assumed to be independent. That is, the joint distribution P_U is a product distribution.

Example 2.6 (Continuation of Example 2.1) Specification of an SCM
In the following, we give an example of how an SCM could be explicitly specified for the data structure (W, A, Z, Y), where W is just a single continuous covariate. In this example, we will assume that we have complete knowledge on the distribution of U, W, A and Z. Let us start by specifying the joint distribution P_U of U

$$\begin{aligned} U_W &\sim \mathcal{N}(0, 1) \\ U_A &\sim \text{Bernoulli}(1/2) \\ U_Z &\sim \text{Bernoulli}(1/50) \\ U_Y &\sim \mathcal{N}(0, 1). \end{aligned} \tag{2.2.6}$$

Now, we can specify the structural equations

$$\begin{aligned} W &= 10 + 2 \cdot U_W \\ A &= U_A \\ Z &= U_Z \\ Y &= \alpha \cdot W + \beta \cdot A + \gamma \cdot Z + U_Y, \end{aligned} \tag{2.2.7}$$

where $\alpha, \beta \in \mathbb{R}$ are unknown parameters. Hence, it is easy to see that we only need to find α, β and γ to have complete knowledge of the DGP, which will enable us to answer questions about features of the DGP.

The representation in (2.2.5), shows an example of a SCM that imposes very few assumptions on the relationships among the involved variables, hence very few restrictions on the model. The model is only augmented with the assumption that the intuitive time ordering holds, namely that a change in Y would not be able to change the values of W nor A. However, a change in A might have an effect on Y, which is exactly the effect of interest in most studies. However, as shown in Example 2.6, the SCM can be specified in a way that imposes additional assumptions by specifying the distribution of U and the class of the functions in f.

It is often of interest to visualise the dependencies among the stochastic variables involved in the SCM and hence make some of the imposed assumptions even more explicit. For this purpose we can utilise graphs to make a quick and easy overview of the possible dependencies among the variables, which gives a more intuitive understanding of some of the assumptions that are imposed in the model.

Definition 2.8 (Graph terminology)**.** A *graph* is a collection of vertices and edges connecting them. Graphs are visualised using nodes and lines representing vertices and edges respectively. A *directed graph* is a graph, where the edges have a certain direction represented by arrows. If an arrow points towards a vertex, it is called a *child* and the vertex that the arrow points away from is called the *parent* of the child. The set of parents of a vertex V is denoted $\text{Pa}(V)$.

In this project, we use *directed acyclic graphs* (DAG) to depict causal relationships. In these graphs, each variable involved in the causal relationship is associated with a unique vertex. The interpretation of the parent-child relationship in a DAG is that the parent potentially affects the child in some way. Every SCM can be associated with a graph, which visualises the causal assumptions about variable dependencies that have been made for each of the variables.

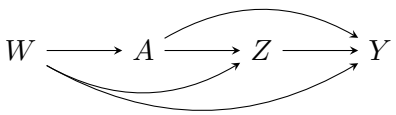


Figure 2.1: *DAG related to* (2.2.5)*. Arrows indicate a potential causal relationship.*

The DAG corresponding to the SCM in (2.2.5) is shown in Figure 2.1. Here, the arrows represent a potential causal relationship. The graph in Figure 2.1 gives a nice overview of the dependencies among variables. However, it is important to understand that there are limitations of DAGs and

that they are not able to encode all assumptions made in the model implied by the SCM, like the functional form of f. Hence, DAGs serve as an addition to the SCM and not something that can stand alone.

Example 2.7 (Continuation of Example 2.1) SCM for an RCT
In a two-armed RCT, we know that the covariate values W do not influence the treatment A and most often we also know the exact form of the mechanism f_A and distribution of U_A. In this case, the SCM would be

$$\begin{aligned} W &= f_W(U_W) \\ A &\sim \text{Bernoulli}(b) \\ Z &= f_Z(W, A, U_Z) \\ Y &= f_Y(W, A, Z, U_Y), \end{aligned} \tag{2.2.8}$$

where $b \in (0, 1)$ is determined by the randomisation ratio. These equations are an example of incorporating the knowledge of the randomisation of treatment in an RCT into the model. As we do not have any knowledge on how the variables are related, f_W, f_Z, f_Y remain unknown functions and likewise the distributions of the errors U_W, U_Z, U_Y also remain unknown.

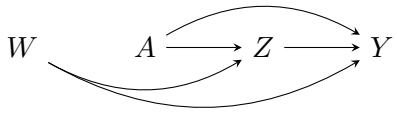


Figure 2.2: *DAG related to* (2.2.8)*. Arrows indicate a potential causal relationship.*

The DAG associated with the SCM (2.2.8) is illustrated in Figure 2.2. As A does not depend on anything but U_A, we will not have any incoming arrows to A in the DAG and as a consequence, we do not have any confounders in the A-Y relationship in an RCT.

Note that in both (2.2.5) and (2.2.8), we could just as well have used the parent-notation introduced in Definition 2.8, and hence written

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(\text{Pa}(A), U_A) \\ Z &= f_Z(\text{Pa}(Z), U_Z) \\ Y &= f_Y(\text{Pa}(Y), U_Y) \end{aligned} \tag{2.2.9}$$

instead, as long as the SCM is associated with a DAG making it clear how the parent-child relations are defined. Hence SCMs and DAGs complement each other very nicely in defining and visualising the data structure and dependencies among the measured variables.

Interventions and g-computation formula

The concept of SCMs and DAGs open up for considering child variables under different distributions of the parent variables, using the terminology introduced in Definition 2.8. This is extremely useful as we are often interested in scenarios where we change the parent variables and see the effect it has in the child variable. If we learn the distribution of the child variable in terms of the parent

variables then we can simply incorporate the newly proposed distribution of the parent variable and find the effect of interest.

Definition 2.9 (Intervention)**.** An *intervention* is an external manipulation of an SCM, where the equation for one or more variables are changed to follow a certain distribution. The variables that are manipulated in this way are referred to as *intervention nodes* and the SCM after incorporating an intervention is called the *post-intervention SCM*.

Imposing an intervention on the treatment variable in the system of equations defined by the SCM describes how data would have looked if the treatment had been assigned in such a way.

Example 2.8 (Continuation of Example 2.7) Post-intervention SCM
Consider the post-intervention SCM

$$\begin{aligned} W &= f_W(U_W) \\ A &= 0 \\ Z &= 0 \\ Y &= f_Y(W, 0, 0, U_Y), \end{aligned} \tag{2.2.10}$$

where we intervened by setting $A = 0$ and $Z = 0$. This SCM will construct the potential outcomes $Y(0, 0)$, since we are assigning all the participants to the control group. Similarly, we could intervene by setting $A = 1$ and $Z = 0$, which would give us the potential outcome $Y(1, 0)$. This is exactly what we need to determine the causal estimand introduced in Example 2.3.

Definition 2.10 (Static intervention)**.** An intervention that imposes a constant function for the intervention node is called a *static intervention*.

The static intervention is the simplest type of intervention and it will be the only type that we will consider in this project. As an example of a static intervention, we have imposed a static intervention for both A and Z in (2.2.10), which impacts the distribution of its child node Y.

If we considered the density for $P \in \mathcal{M}$, we could factorise the density p as

$$p(O) = p(W, A, Z, Y) = \underbrace{p(A \mid W)p(Z \mid A, W)}_{g}\underbrace{p(W)p(Y \mid W, A, Z)}_{q}, \tag{2.2.11}$$

where g represents the part of the density corresponding to the intervention nodes A and Z. Note the clear connection between this decomposition and the SCM (2.2.5), as given the functions $f = (f_W, f_A, f_Z, f_Y)$ and the distribution P_U we know (2.2.11). Now, if we replace g with an intervention in (2.2.11), we get the *g-computation formula*, which is the density under the external manipulation. Under the static intervention, as presented in Example 2.8, the densities for A and Z is just indicators. Hence the joint density simplifies to

$$\begin{aligned} p_{\text{int}}(O) &= I(A = 0)I(Z = 0)p(W)p(Y \mid W, A, Z) \\ &= p(W)p(Y \mid W, A = 0, Z = 0). \end{aligned} \tag{2.2.12}$$

2.2.4 Identification

To conduct meaningful causal inference, we require that there is a direct connection between observations in the causal model to observations in the statistical one. This connection between the causal and statistical model relies on non-testable assumptions. For example, if we are interested in the causal estimand introduced in Example 2.3 which involves two of the potential outcomes, making it a feature of the probability distributions in the causal model, we need the following assumptions to be satisfied.

Assumption 2.11 (Identifying assumptions)**.**

(i) **No unmeasured confounders**:
- $[(Y(0,0), Y(0,1), Y(1,0), Y(1,1)) \perp\!\!\!\perp A] \mid W$
- $[(Y(0,0), Y(0,1), Y(1,0), Y(1,1)) \perp\!\!\!\perp Z] \mid A, W$

(ii) **Positivity**:
- $P_0(A = a \mid W) > 0$ for $a \in \{0, 1\}$, whenever $P_0(W) > 0$
- $P_0(Z = z \mid A, W) > 0$, for $z \in \{0, 1\}$, whenever $P_0(A, W) > 0$

(iii) **Consistency**: $Y = Y(a, z)$ whenever $A = a$ and $Z = z$

Assumption 2.11 (i) states that all possible confounders in the A-Y relationship and in the Z-Y relationship should be measured. In other words, the first of these states that we assume conditional independence between the treatment mechanism and the potential outcomes given the covariates. The positivity assumption 2.11 (ii) ensures that any participant in the study has a non-zero probability of getting allocated to either treatment option, for both A and Z. The consistency assumption 2.11 (iii) states that the actual outcome observed in the trial corresponds to the potential outcome determined by the actual treatment assignments. If we draw inference on a causal quantity from a statistical quantity without satisfying the identifying assumptions, it could lead to bias called *identification bias*.

If both A and Z are randomised, assumptions 2.11 (i) and 2.11 (ii) are satisfied by definition. As mentioned, the random assignment of treatment ensures no confounders at all, and hence also no unmeasured confounders. In addition, the randomisation also ensures positive probability of both treatment options regardless of the covariate values. At first sight it might seem natural that Assumption 2.11 (iii) is satisfied in this context as it just says that the outcome you observe is exactly the outcome you thought you would observe. One way this assumption could be violated is, if the treatment group was not uniquely defined. For example, you can not pool observations from patients that receive different doses of treatment into one treatment arm.

The importance of each of these assumptions becomes more clear when showing the identifiability, which is the topic of the following result. *Identifiability* refers to having a connection between a feature of distributions in the causal model and a feature of distributions in the statistical model.

Proposition 2.12 (Identifiability)**.** *Let $P^* \in \mathcal{M}^*$ denote the true DGP for $O^* = (W, A, Z, Y(0,0), Y(0,1), Y(1,0), Y(1,1))$ and $P_0 \in \mathcal{M}$ the true DGP for $O = (W, A, Z, Y)$. Then under Assumption 2.11, it holds that*

$$\begin{aligned} E_{P^*}[Y(1,0) - Y(0,0)] = E_{P_0}\big[&E_{P_0}[Y \mid A = 1, Z = 0, W] \\ &- E_{P_0}[Y \mid A = 0, Z = 0, W]\big], \end{aligned} \tag{2.2.13}$$

where the subscript in the expectation makes it explicit that the expectation is taken over the

distribution P^* or P_0.

Proof. By the law of total expectation we find that

$$\begin{aligned} E_{P^*}[Y(1,0) - Y(0,0)] &= E_{P^*}[E_{P^*}[Y(1,0) - Y(0,0) \mid W]] \\ &= E_{P^*}[E_{P^*}[Y(1,0) \mid A = 1, W]] \\ &\quad - E_{P^*}[E_{P^*}[Y(0,0) \mid A = 0, W]] \\ &= E_{P^*}[E_{P^*}[Y(1,0) \mid A = 1, Z = 0, W]] \\ &\quad - E_{P^*}[E_{P^*}[Y(0,0) \mid A = 0, Z = 0, W]] \\ &= E_{P_0}[E_{P_0}[Y \mid A = 1, Z = 0, W] \\ &\quad - E_{P_0}[Y \mid A = 0, Z = 0, W]], \end{aligned} \tag{2.2.14}$$

where the second and third equality utilises the no unmeasured confounders assumption 2.11 (i) and linearity of the expectation, and the fourth uses the assumption of consistency 2.11 (iii). The positivity assumption 2.11 (ii) ensures that these expectations are well-defined. □

The result in Proposition 2.12 shows that Assumption 2.11 ensures that we can identify the causal estimand introduced in Example 2.3, a quantity in the causal model, from a quantity in the statistical model, which is exactly equal to the one in Example 2.5. This allows us to make causal inference and make a causal interpretation from statistical inference of observable data. Hence it is extremely important what assumptions we impose, as it will be crucial for the identification result.

Note that the assumptions in 2.11 are stricter than what is actually used to show the identification in Proposition 2.12. As an example, we only consider the potential outcomes for which $Z = 0$ and hence we can actually weaken the positivity assumption on this variable to be $P_0(Z = 0 \mid A, W) > 0$ whenever $P_0(A, W) > 0$.

In this chapter, we have given a description of the data structure, which led to defining models and estimands in the causal and statistical worlds in Figure 1.1. In addition, we established a connection between these worlds in terms of identification. Hence the statistical estimation problem is now fully defined. Next up is the actual estimation part, which will be the topic of the next chapter, before we are able to make meaningful causal inference.

3 Semiparametric theory

In Chapter 1 and 2, we presented a method for properly defining the clinical research question at hand, and translating it to a statistical estimand. This chapter aims to elucidate what is considered a meaningful way to answer the clinical question based on data.That is, we are interested in finding estimators of the quantity of interest that have certain properties that make the estimator, and hence the estimate, sensible. The chapter is based on [12], [16], [17], [18, App. B.1] and [19].

Previously, we introduced the concept of statistical models. The most well-known class of models is *parametric models*, where the densities can be described by a finite dimensional parameter θ. This could for example be Gaussian distributions, indexed by the two-dimensional parameter $\theta = (\mu, \sigma)$. Parametric models can be quite restrictive. Often, one does not believe that the DGP is parametric, however, with the words "all models are wrong, but some are useful" of George E.P. Box in mind, various parametric models are often fitted and interpreted anyway [20]. Of course parametric models are favourable as one can use maximum likelihood estimation to obtain an estimate and make valid inference under regularity conditions. But the choice of parametric model is often influenced by the researcher's preferences, the type of outcome being measured and even distributions in the data, with limited consideration about the underlying DGP. As a consequence, when restricting the focus to parametric models it is likely that the statistical model no longer contains the true DGP and hence we will never reach the truth inside this model, at most we are able to reach an approximation. Therefore, it is important that the statistical model is built on factual knowledge of the DGP.

Based on these considerations, we may want to consider models that are not restricted to densities that can be described by a finite-dimensional parameter, that is, look towards models that contain distributions that depend on an infinite dimensional parameter. In the case where we make no restrictions on the infinite dimensional parameter, denoted by θ, the model will be referred to as a *nonparametric model*. Nonparametric models are often considered when little to nothing is known about the underlying data-generating mechanisms. These models offer flexibility by not assuming a specific functional form, making them suitable for capturing complex and unknown relationships between variables. Hence, they allow for realistically modelling the complexity of the true DGP, however this adaptivity also makes estimation and inference more difficult.

In an RCT, the treatment mechanism is known a priori, hence we know that part of the real-world distribution can actually be described by a finite dimensional parameter. We can implement this knowledge into the nonparametric model, which results in a *semiparametric model* where the parameter is made up by a finite and an infinite dimensional part. That is, the infinite dimensional parameter θ has a finite dimensional parameter component. Semiparametric models offer a balance between incorporating our knowledge of how the experiment was generated and allowing flexibility to capture potential complex relationships in the true DGP.

3.1 Estimators

In Subsection 2.2.4 we established identifiability of an unobservable causal parameter by a statistical parameter, under certain assumptions. This statistical parameter is a feature of an observable distribution. We have constructed a statistical estimation problem which, with the right tools, is a manageable task. In this chapter, we will move to the next part of the workflow Figure 1.1, which is the estimator.

We will focus on one-dimensional features of the true DGP. An *estimate* is an approximation of the true value of this feature, based on observed data. However, it is of interest that this estimate depends on the observed data in such a way that we end up making a qualified approximation of the quantity. What is exactly meant by a qualified approximation will be the main topic of this section.

We start by introducing some notation that will be used throughout the chapter. Let $\mathcal{M}$ be a statistical model for some finite-dimensional stochastic vector O that follows an unknown probability distribution P_0. Let $o_1, \ldots, o_n$ denote n i.i.d. observations of the stochastic variable O. In addition we will use $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(O_i)$ as a shorthand notation for empirical mean of a function evaluated in the n i.i.d. stochastic variables $O_1, \ldots, O_n$.

Definition 3.1 (Estimator)**.** Let $o_1, \ldots, o_n$ denote n i.i.d. realisations of the stochastic variable $O \sim P_0 \in \mathcal{M}$, where $\mathcal{M}$ is a statistical model. An estimator is a function of i.i.d. stochastic vectors:

$$\hat{\Psi}_n : \mathcal{O}^n \to \mathbb{R}, \tag{3.1.1}$$

where $\mathcal{O}$ is the sample space of O and $\hat{\Psi}_n(o_1, \ldots, o_n)$ is interpreted as an estimate of the true value $\Psi(P_0)$.

Often we will write $\hat{\Psi}_n$ as shorthand for the estimator or estimate and it should be clear from context what is meant.

Example 3.1
Consider a situation where we are interested in the mean of some one-dimensional variable $X \sim P_0$ in a population. In this case the sample mean $\hat{\Psi}_n(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an estimator for the population mean $\mu = E_{P_0}[X]$. Given observed data $(x_1, \ldots, x_n)$, our estimator creates an estimate $\hat{\mu}_n = \hat{\Psi}_n(x_1, \ldots, x_n)$ which in this case is just a scalar.

It is not hard to imagine that there are many possibilities for defining an estimator, and hence we are interested in ways to determine if the estimator returns good estimates. Of course, this requires us to define what is meant by a good estimate. First, we are interested in building an estimator that, with enough data, converges to the true parameter value, which is known as consistency.

Definition 3.2 (Consistent estimator)**.** Consider a sequence of estimators $\left(\hat{\Psi}_n\right)_{n\in\mathbb{N}}$, where $\hat{\Psi}_n : \mathcal{O}^n \to \mathbb{R}$. $\hat{\Psi}_n$ is said to be a *consistent estimator* if

$$\hat{\Psi}_n(O_1, \ldots, O_n) \xrightarrow{P_0} \Psi(P_0), \quad n \to \infty. \tag{3.1.2}$$

In Definition 3.2, $\xrightarrow{P_0}$ denotes convergence in probability with respect to P_0, for a formal definition see Definition A.3.

Moreover, we are interested in an estimator that has a known sampling distribution asymptotically such that we can construct valid confidence intervals and make inference. In addition, it would be preferred that this distribution has the lowest possible variance to make the confidence intervals as tight as possible. These properties will be discussed in greater detail in the following subsections.

3.1.1 Asymptotic linearity

We want an estimator that turns observations into a good estimate of the true parameter. However, to be more clear about what makes an estimator good, we will introduce the concept of asymptotic linearity. This ensures, that for large samples, the estimator converges to the right answer in probability and converges in distribution to a sampling distribution that allows us to do inference.

Definition 3.3 (Asymptotically linear estimator)**.** Consider an estimator $\hat{\Psi}_n : \mathcal{O}^n \to \mathbb{R}$ of the statistical estimand $\Psi : \mathcal{M} \to \mathbb{R}$. If there exists a function $\phi_{P_0} : \mathcal{O} \to \mathbb{R}$ that satisfies $\phi_{P_0} \in \mathcal{L}_0^2(P_0)$ such that

$$\hat{\Psi}_n - \Psi(P_0) = \mathbb{P}_n \phi_{P_0} + o_{P_0}\left(n^{-1/2}\right), \tag{3.1.3}$$

we say that the estimator $\hat{\Psi}$ is *asymptotically linear*.

The function $\phi_{P_0} \in \mathcal{L}_0^2(P_0)$ that satisfies (3.1.3) for some estimator will be referred to as the *influence function* (IF) corresponding to the estimator. The IF is a function of the stochastic variable O that depends on the unknown distribution P_0 and it measures the sensitivity of an estimator to small changes or perturbations in the data [21]. The space $\mathcal{L}_0^2(P_0)$, in which the IF is contained, is the Hilbert space with functions that have finite variance and mean zero under P_0, see Definition A.10 in Appendix A.

The notation $V_n = o_{P_0}(n^{-1/2})$ is used to express the rate of convergence in probability of a sequence $(V_n)_{n\in\mathbb{N}}$, meaning that

$$V_n = o_{P_0}(n^{-1/2}) \quad \Longrightarrow \quad \sqrt{n}V_n \xrightarrow{P_0} 0, \text{ for } n \to \infty. \tag{3.1.4}$$

Example 3.2
We will illustrate the derivation of an IF by a simple example. Consider a univariate stochastic variable $O \sim P_0$ with finite variance. We will consider a nonparametric statistical model $\mathcal{M}$ and the statistical estimand $\Psi : P \in \mathcal{M} \mapsto E_P[O]$. Consider the estimator $\hat{\Psi}_n = \mathbb{P}_n O$, which is simply the sample mean. The estimator $\hat{\Psi}_n$ is an asymptotically linear estimator of $\Psi(P_0)$ as

$$\hat{\Psi}_n - \Psi(P_0) = \mathbb{P}_n\left(O - \Psi(P_0)\right) \tag{3.1.5}$$

which shows that the IF for this estimator is

$$\phi_{P_0}(O) = O - \Psi(P_0) = O - E_{P_0}[O]. \tag{3.1.6}$$

We can check that the IF is in $\mathcal{L}_0^2(P_0)$ by checking that it has mean zero

$$E_{P_0}[\phi_{P_0}(O)] = E_{P_0}[O] - E_{P_0}[O] = 0 \tag{3.1.7}$$

and finite variance

$$\text{Var}_{P_0}[\phi_{P_0}(O)] = \text{Var}_{P_0}[O] < \infty. \tag{3.1.8}$$

It is unfortunately not always as easy to derive the IF of an asymptotically linear estimator as we saw in Example 3.2. However, as there is a lot of literature on the topic of deriving IFs, we will not go much more in detail with this and instead refer to derivations of IFs when needed [17].

Proposition 3.4. *An asymptotically linear estimator $\hat{\Psi}_n$ of $\Psi(P_0)$, with influence function ϕ_{P_0}, satisfies*

$$\sqrt{n}\left(\hat{\Psi}_n - \Psi(P_0)\right) \xrightarrow[P_0]{d} \mathcal{N}(0, \sigma_0^2), \tag{3.1.9}$$

for $n \to \infty$.

Proof. By the central limit theorem (CLT) A.5, it holds that

$$\sqrt{n}\mathbb{P}_n\phi_{P_0} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_{P_0}(O_i) \xrightarrow[P_0]{d} \mathcal{N}(0, \sigma_0^2) \tag{3.1.10}$$

as $E_{P_0}[\phi_{P_0}] = 0$ and σ_0^2 denotes the variance of the IF ϕ_{P_0}, which is referred to as the *asymptotic variance*. Combining this with (3.1.3) using Slutsky's theorem A.4 we know that an asymptotically linear estimator satisfies

$$\sqrt{n}\left(\hat{\Psi}_n - \Psi(P_0)\right) \xrightarrow[P_0]{d} \mathcal{N}(0, \sigma_0^2). \tag{3.1.11}$$

□

By this result, asymptotic linearity guarantees that the estimator is consistent, see Definition 3.2. Proposition 3.4 is extremely useful as it implies that we can find the asymptotic distribution of the estimator by knowing the IF. The asymptotic distribution of the estimator is important for constructing confidence intervals and doing inference. Denoting the estimate of σ_0^2 by $\hat{\sigma}_0^2$ and specifying a level of significance α, we may obtain an approximate $(1-\alpha)\cdot 100\%$ confidence interval for $\Psi(P_0)$ by (3.1.11):

$$\hat{\Psi}_n \pm z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_0^2}{n}}, \tag{3.1.12}$$

where $z_{\alpha/2}$ denotes the $1-\alpha/2$-quantile of the standard normal distribution.

Theorem 3.5 ([16])**.** *An asymptotically linear estimator has an influence function ϕ that is almost surely unique.*

Proof. We will establish this result through a proof by contradiction.
Suppose the contrary. Then, there must exist an alternative IF ϕ^* such that $E[\phi^*] = 0$, and

$$\sqrt{n}(\hat{\Psi}_n - \Psi(P_0)) = \sqrt{n}\mathbb{P}_n\phi^* + o_{P_0}(1). \tag{3.1.13}$$

Since ϕ is also an IF, then $\sqrt{n}(\hat{\Psi}_n - \Psi(P_0))$ is also equal to $\sqrt{n}\mathbb{P}_n\phi + o_{P_0}(1)$, this implies that

$$\sqrt{n}\mathbb{P}_n(\phi - \phi^*) = o_{P_0}(1). \tag{3.1.14}$$

We can apply the CLT, Theorem A.5, to obtain

$$\sqrt{n}\mathbb{P}_n(\phi - \phi^*) \xrightarrow[P_0]{d} \mathcal{N}\left(0, E\left[(\phi - \phi^*)^2\right]\right), \tag{3.1.15}$$

where Var$[\phi - \phi^*] = E\left[(\phi - \phi^*)^2\right]$ since it is a mean zero random variable. For this limiting normal distribution to be $o_{P_0}(1)$, it necessitates that the covariance matrix satisfies

$$E\left[(\phi - \phi^*)^2\right] = 0^{q\times q}, \tag{3.1.16}$$

which then implies that $\phi = \phi^*$ almost surely. □

3.1.2 Regularity

As seen in (3.1.11), asymptotic linearity of an estimator ensures that we have a known asymptotic distribution when estimating the target parameter. This is a property that ensures nice behaviour of the estimator. However, we also want some type of control as we go towards the true distribution P_0 and the local area around P_0. It is of course favourable that the estimator has some type of robustness such that we do not end up with very different estimates for very similar distributions.

In short, regular estimators are estimators that have a limiting distribution that does not change when local changes are made to the underlying DGP. Before giving a rigorous definition we need to introduce the concept of a path and its corresponding score. This will serve as a technical device for analysing semiparametric models and constructing tangent spaces, which will be defined later in this chapter.

Definition 3.6 (Path)**.** For an arbitrary distribution $P \in \mathcal{M}$ we define a *path* through P as $\left\{\tilde{P}_\varepsilon : \varepsilon \in \mathbb{R}\right\} \subseteq \mathcal{M}$ which satisfies $\tilde{P}_\varepsilon|_{\varepsilon=0} = P$ and has score function $h \in \mathcal{L}_0^2(P)$ equal to

$$h = \frac{d}{d\varepsilon} \log \tilde{p}_\varepsilon \Big|_{\varepsilon=0}, \tag{3.1.17}$$

where $\tilde{p}_\varepsilon$ denotes the density of the distribution $\tilde{P}_\varepsilon$.

The paths that we will consider in this project are assumed to satisfy certain smoothness conditions, which is a technicality that we will not go into in this project, for more see [16].

By this definition it is not surprising that in some literature, a path is also referred to as a parametric submodel, since it is parametrised by a one-dimensional parameter ε and is contained in the statistical model $\mathcal{M}$. However, a path is not a parametric model in the usual sense. It should be considered as a purely theoretic tool for generalizing methods known from parametric model theory to semi- and nonparametric models. The following example will illustrate one way to formulate distributions along a path and some properties of the corresponding score.

Example 3.3
Let $\mathcal{M}$ be a nonparametric statistical model. Consider $P, \tilde{P} \in \mathcal{M}$ and let p and $\tilde{p}$ denote

their respective density functions. For $\varepsilon \in [0,1]$ define $\tilde{P}_\varepsilon$ by having density

$$\tilde{p}_\varepsilon(o) = \varepsilon\tilde{p}(o) + (1-\varepsilon)p(o). \tag{3.1.18}$$

Note that for each $\varepsilon \in [0,1]$ $\tilde{p}_\varepsilon(o)$ is a valid density by definition since it is non-negative and

$$\int \tilde{p}_\varepsilon(o)do = \varepsilon \int \tilde{p}(o)do + (1-\varepsilon)\int p(o)do = 1. \tag{3.1.19}$$

Then the set of distributions $\left\{\tilde{P}_\varepsilon : \varepsilon \in [0,1]\right\} \subseteq \mathcal{M}$ define a path from P towards $\tilde{P}$. It is important that the model $\mathcal{M}$ is assumed nonparametric in order to ensure that this path between two arbitrary distributions in the model is also contained in the model. Had the model instead been semiparametric, one would need to be careful when defining the path such that the distributions along the path are included in the model. Since the score function is just the change in the log-likelihood at P, it can be expressed as

$$\begin{aligned} h(o) &= \frac{d}{d\varepsilon}\log\left(\tilde{p}_\varepsilon(o)\right)\Big|_{\varepsilon=0} = \frac{d}{d\varepsilon}\log\left(\varepsilon\tilde{p}(o) + (1-\varepsilon)p(o)\right)\Big|_{\varepsilon=0} \\ &= \frac{\tilde{p}(o) - p(o)}{\varepsilon\tilde{p}(o) + (1-\varepsilon)p(o)}\Bigg|_{\varepsilon=0} = \frac{\tilde{p}(o)}{p(o)} - 1. \end{aligned} \tag{3.1.20}$$

We are able to write paths with densities in terms of the score h, which gives a sense of direction, instead of using the destination $\tilde{P}$. By rewriting (3.1.18), using (3.1.20), we obtain

$$\begin{aligned} \tilde{p}_\varepsilon(o) &= \varepsilon\tilde{p}(o) + (1-\varepsilon)p(o) \\ &= \varepsilon(h(o)+1)p(o) + (1-\varepsilon)p(o) \\ &= \varepsilon h(o)p(o) + \varepsilon p(o) + p(o) - \varepsilon p(o) \\ &= (1+\varepsilon h(o))p(o), \end{aligned} \tag{3.1.21}$$

which depends on the direction h rather than the destination $\tilde{P}$.

In Example 3.3, we have shown that we can easily write the score from a path in terms of the density functions of the distributions that the path starts and ends in. Now, we will show that given a zero mean function with finite variance, we can construct a path with score equal to a given function in $\mathcal{L}_0^2(P)$.

Proposition 3.7. *Consider a function $h \in \mathcal{L}_0^2(P)$ such that $h(o) \geq -1$ for all $o \in \mathcal{O}$. A path $\left\{\tilde{P}_\varepsilon : \varepsilon \in [0,1]\right\} \subseteq \mathcal{M}$ starting in P with score h can be constructed by letting $\tilde{P}_\varepsilon$ be the distribution with density*

$$\tilde{p}_\varepsilon(o) = (1+\varepsilon h(o))p(o). \tag{3.1.22}$$

Proof. First we need to show that $\tilde{p}_\varepsilon$ is a density. As $h(o) \geq -1$ it holds that

$$\tilde{p}_\varepsilon = (1+\varepsilon h(o))p(o) \geq 0 \tag{3.1.23}$$

for all $\varepsilon \in [0,1]$. In addition, it is quite easy to see that

$$\begin{aligned}\int \tilde{p}_\varepsilon(o)do &= \int (1+\varepsilon h(o))p(o)do = \int p(o)do + \varepsilon \int h(o)p(o)do \\ &= 1 + \varepsilon E_P[h(O)] = 1\end{aligned} \tag{3.1.24}$$

as $E_P[h(O)] = 0$. Second, we need to show that the resulting distribution $\tilde{P}_\varepsilon$ is contained in the model for $\varepsilon \in [0,1]$. If $\mathcal{M}$ is assumed nonparametric we can move freely in the model space no matter the starting point ensuring $\tilde{P}_\varepsilon \in \mathcal{M}$ for all $\varepsilon \in [0,1]$. However, if $\mathcal{M}$ is assumed semiparametric, then there might a concern about whether or not the resulting distributions are contained in the model. To avoid this problem we will assume that there always exists a $\mathcal{E} \in \mathbb{R}$ such that $\varepsilon < \mathcal{E}$ implies $\left\{\tilde{P}_\varepsilon : \varepsilon \in [0, \min(1,\mathcal{E})]\right\} \subseteq \mathcal{M}$ [12].

Now it remains to show that the path actually has score h. Using (3.1.17), we find that

$$\begin{aligned}\frac{d}{d\varepsilon}\log \tilde{p}_\varepsilon(o)\,\Big|_{\varepsilon=0} &= \frac{d}{d\varepsilon}\log((1+\varepsilon h(o))p(o))\,\Big|_{\varepsilon=0} \\ &= \frac{h(o)p(o)}{(1+\varepsilon h(o))p(o)}\,\Big|_{\varepsilon=0} \\ &= \frac{h(o)p(o)}{p(o)} = h(o),\end{aligned} \tag{3.1.25}$$

and hence $\left\{\tilde{P}_\varepsilon : \varepsilon \in [0,1]\right\} \subseteq \mathcal{M}$ is a path with score h. □

Returning to the characteristic of a regular estimator, the following definition will define when an estimator is considered regular.

Definition 3.8 (Regular estimator)**.** Consider any sequence of probability distributions $\left(\tilde{P}_{n^{-1/2}}\right)_{n\in\mathbb{N}}$ along any path through P_0, where $\tilde{P}_{n^{-1/2}}$ goes toward the truth P_0 as $n \to \infty$. Let $\left(\hat{\tilde{\Psi}}_n\right)_{n\in\mathbb{N}}$ be a sequence of estimators, where $\hat{\tilde{\Psi}}_n$ denotes an estimate of $\Psi(\tilde{P}_{n^{-1/2}})$ given n observations from $\tilde{P}_{n^{-1/2}}$. The estimator is *regular* at P_0 for estimating $\Psi(P_0)$ if the sequence $\left(\hat{\tilde{\Psi}}_n\right)_{n\in\mathbb{N}}$ satisfies

$$\sqrt{n}\left(\hat{\tilde{\Psi}}_n - \Psi(\tilde{P}_{n^{-1/2}})\right) \xrightarrow[\tilde{P}_{n^{-1/2}}]{d} \mathcal{D}_{P_0}, \tag{3.1.26}$$

where $\mathcal{D}_{P_0}$ is a known distribution that is independent of the sequence $\left(\tilde{P}_{n^{-1/2}}\right)_{n\in\mathbb{N}}$.

Regularity ensures some kind of smoothness of the mapping $\hat{\Psi}_n$, and that it is robust to small perturbations in the underlying DGP. It is important to notice that since we are considering the sequence of probability distributions, the convergence in distribution is with respect to different distributions as we go through the sequence. But, as n gets large, the difference in the distributions becomes smaller. The following example will make it more explicit, what is actually meant by this convergence in distribution, when the distribution changes.

Example 3.4

Consider a sequence of estimators $\left(\hat{\tilde{\Psi}}_n\right)_{n\in\mathbb{N}}$. By Lemma A.2, the convergence in distribution presented in (3.1.26) implies, for $Z \sim \mathcal{D}_{P_0}$ and any continuous and bounded function f, that

$$E\big[f\big(\sqrt{n}(\hat{\Psi}_n(O_{1,n},\ldots,O_{n,n}) - \Psi(\tilde{P}_{n^{-1/2}}))\big) - f(Z)\big] \to 0 \text{ as } n \to \infty. \tag{3.1.27}$$

For each n, we have that $O_{1,n},\ldots,O_{n,n}$ are n independent vectors that follow the distribution $\tilde{P}_{n^{-1/2}}$. That is, $O_{i,j}$ is the i'th vector that follows the distribution $\tilde{P}_{j^{-1/2}}$, which means that for $n=1$ we only have a single vector $O_{1,1}$ and for $n=2$ we have $O_{1,2}, O_{2,2}$ and so on. As n changes, the distribution changes, the vectors used in the estimator change and $\Psi(\tilde{P}_{n^{-1/2}})$ changes too.

Estimators that are not regular can behave badly outside of a very small and specific set of distributions, hence it makes sense to restrict our focus on regular estimators. Under regularity conditions, some of the well-known estimators are regular. For example the sample mean, the MLE and the OLS are regular estimators. Regular estimators possesses desirable asymptotic properties that allow for valid statistical inference as the sample size becomes large.

Consider a regular and asymptotically linear (RAL) estimator. As $\mathcal{D}_{P_0}$ does not depend on the sequence $(\tilde{P}_{n^{-1/2}})_{n\in\mathbb{N}}$ in Definition 3.8 we can choose the trivial sequence $\tilde{P}_{n^{-1/2}} := P_0$ for $n \in \mathbb{N}$ in (3.1.26) to conclude that $\mathcal{D}_{P_0} = \mathcal{N}(0,\sigma_0^2)$ by asymptotic linearity (3.1.11). Hence, if an estimator is RAL, it means that as we move towards the truth, we still tend to the same normal distribution no matter the chosen sequence $\left(\tilde{P}_{n^{-1/2}}\right)_{n\in\mathbb{N}}$ along a path through P_0, that is

$$\sqrt{n}\left(\hat{\tilde{\Psi}}_n - \Psi(\tilde{P}_{n^{-1/2}})\right) \xrightarrow[\tilde{P}_{n^{-1/2}}]{d} \mathcal{N}\left(0, E_{P_0}[\phi_{P_0}^2]\right). \tag{3.1.28}$$

In addition, it can be shown that the regular estimator with the smallest variance is guaranteed to be an asymptotically linear estimator [22, Theorem 25.20]. Therefore, it makes sense to limit ourselves to RAL estimators going forward.

3.1.3 Asymptotic efficiency

It is easy to argue that asymptotic linearity and regularity are attractive properties. However, in many cases there will be multiple RAL estimators of the estimand of interest. In this case, one might question, which one do we choose? To answer this question, we introduce the concept of efficiency. With enough data, the only difference between different RAL estimators is the asymptotic variance imposed by the variance of their respective IFs. When doing inference it is of course preferable that the confidence intervals (3.1.12) are as small as possible, while still being valid. This means that we favour estimators where the asymptotic variance is as small as possible.

Definition 3.9. Let $P_0 \in \mathcal{M}$, where $\mathcal{M}$ is a statistical model. Furthermore, let $\Psi : \mathcal{M} \to \mathbb{R}$ be a statistical estimand. A RAL estimator $\hat{\Psi}$ of Ψ, that has IF $\phi^*_{P_0}$, is called *asymptotically efficient* if

$$\text{Var}[\phi^*_{P_0}] \leq \text{Var}[\phi_{P_0}] \tag{3.1.29}$$

for any other IF ϕ_{P_0} of another RAL estimator of the same statistical estimand.

The IF $\phi^*_{P_0}$ satisfying (3.1.29) is called the *efficient influence function* (EIF) for the statistical estimand. Derivation of EIFs is well-documented in the existing literature for the most common statistical estimands, hence we will refer to [17] for derivation. The following example will present an EIF for the estimation problem that was introduced in Example 2.5.

Example 3.5 (Continuation of Example 2.5)
Consider the statistical estimation problem in Example 2.5 but where we disregard the additional treatment variable Z and hence consider the case with a single treatment variable, resulting in the data structure $O = (W, A, Y)$. The statistical estimand can be expressed as a difference of two other statistical estimands where

$$\Psi_0(P) = E_P\left[E_P[Y \mid A = 1, W]\right] \tag{3.1.30}$$
$$\Psi_1(P) = E_P\left[E_P[Y \mid A = 0, W]\right], \tag{3.1.31}$$

making $\Psi(P) = \Psi_1(P) - \Psi_0(P)$. Sometimes it is easier to consider the statistical estimands Ψ_1 and Ψ_0 separately. By [12, sec. 3.3], it turns out that the EIF ϕ^*_P of Ψ is simply the difference of the EIFs $\phi^*_{1,P}$ and $\phi^*_{0,P}$ corresponding to Ψ_1 and Ψ_0 respectively.
Let $\bar{Q}(A, W) = E_P[Y \mid A, Z, W]$ and $\bar{G}(W) = E_P[A \mid W] = E_P[A]$. For $P \in \mathcal{M}$, the EIF ϕ^*_P is given by

$$\begin{aligned}\phi^*_P &= \phi^*_{1,P} - \phi^*_{0,P} \\ &= \left(\frac{I(A=1)}{\bar{G}(W)} - \frac{I(A=0)}{1-\bar{G}(W)}\right)(Y - \bar{Q}(A,W)) \\ &\quad + \bar{Q}(1,W) - \bar{Q}(0,W) - \Psi(P).\end{aligned} \tag{3.1.32}$$

In the following example, we will give a direct application of how to estimate the EIF in a realisation of the stochastic variable $O = (W, A, Y)$ after observing data and making an initial estimate of the target parameter.

Example 3.6 (Continuation of Example 3.5)
Consider an estimate $\hat{P}$ of P_0 and assume that $\hat{Q}_Y$, the conditional distribution of Y under $\hat{P}$, is given by $Y \mid W, A \sim \mathcal{N}(0.95W - 0.3A - 0.1A \cdot W, 1)$, with $E_{\hat{P}}[A] = 0.5$ and that the initial estimated effect is -0.42. For a participant where we have observed $o_1 = (8.7, 1, 8.2)$, the EIF in o_1 is

$$\hat{\phi}^*_{\hat{P}}(o_1) = \left(\frac{1}{0.5} - \frac{0}{0.5}\right)(8.2 - 7.095) + 7.095 - 8.265 + 0.42 = 1.46, \tag{3.1.33}$$

by (3.1.32). It is clear to see, that an observation which deviates more from the general tendency in data, will end up with a larger value of the EIF.

3.2 Characterising the set of influence functions

We have seen in the previous sections, that it is of interest to find the EIF as this leads to an efficient RAL estimator for which we know the asymptotic distribution. This enables us to construct valid

confidence intervals and make inference from the observed data.

Definition 3.10 (Gâteaux derivative, pathwise differentiable, gradient)**.** Let $\left\{\tilde{P}_\varepsilon : \varepsilon \in [0,1]\right\}$ be a path starting in $P \in \mathcal{M}$ and let Ψ be a statistical estimand. When the limit

$$\lim_{\varepsilon \to 0} \frac{\Psi(\tilde{P}_\varepsilon) - \Psi(P)}{\varepsilon} \tag{3.2.1}$$

exists, it is called the *Gâteaux derivative* of Ψ in P and will be denoted by $\nabla_h \Psi(P)$. Additionally, Ψ is called *pathwise differentiable* at P if there exists a function $\phi_P : \mathcal{O} \to \mathbb{R}$ such that for every path $\left\{\tilde{P}_\varepsilon : \varepsilon \in [0,1], \tilde{P}_\varepsilon|_{\varepsilon=0} = P\right\} \subset \mathcal{M}$ with score h, it holds that

$$\nabla_h \Psi(P) = E_P[h\phi_P]. \tag{3.2.2}$$

A function ϕ_P that satisfies (3.2.2) is called a *gradient* of Ψ. The term on the right hand side is the inner product of the score h and function ϕ_P in $\mathcal{L}_2^0(P)$, and it tells us the angle between the score of the path and the IF of the estimator. On the other hand, the Gâteaux derivative describe the changes in the estimand as we move along the path with score h.

Theorem 3.11 ([18])**.** *Let $\Psi : \mathcal{M} \to \mathbb{R}$ be a statistical estimand. Under certain regularity conditions, if a function $\phi_P \in \mathcal{L}_0^2(P)$ satisfies* (3.2.2)*, then there exists a RAL estimator of $\Psi(P)$ with influence function ϕ_P.*

There are multiple ways of building a RAL estimator from a function $\phi_P \in \mathcal{L}_0^2(P)$ satisfying (3.2.2). In Section 3.3 we will show one way of doing so. In Chapter 4 another method for constructing such an estimator is described. The following theorem will explain why the notation for gradients is exactly the same as for influence functions and state a condition for which some statistical estimands that are pathwise differentiable.

Theorem 3.12 (An influence function is a gradient)**.** *Let $\mathcal{M}$ be a statistical model, $P \in \mathcal{M}$ and $\Psi : \mathcal{M} \to \mathbb{R}$ be the statistical estimand. If one can construct a RAL estimator of $\Psi(P)$, with IF ϕ_P, then Ψ is pathwise differentiable and ϕ_P is a gradient of Ψ at P.*

Since this proof relies on some rather technical results, it is included in Appendix A.3 for better readability here.

Corollary 3.13. *Consider two RAL estimators for the same estimand with IF ϕ_1 and ϕ_2 respectively. The difference between the IFs is orthogonal to any score h in $\mathcal{L}_0^2(P)$ corresponding to paths $\{\tilde{P}_\varepsilon : \varepsilon \in \mathbb{R}\}$ in the statistical model $\mathcal{M}$ through P. That is*

$$E_P[(\phi_1 - \phi_2)h] = 0. \tag{3.2.3}$$

Proof. By Theorem 3.12 we know that (3.2.2) holds for any two RAL estimators with IFs ϕ_1 and ϕ_2. First, note that the left hand side of (3.2.2) solely depends on the statistical estimand Ψ, the

true distribution P, the score h and hence the path. Second, the right hand side solely depends on the choice of estimator through the corresponding IF ϕ_P, the true distribution P, the score h and hence the path. The the Gâteaux derivative $\nabla_h \Psi(P)$ remains the same as this does not depend on the IFs, and hence we can write

$$E_P[h\phi_1] = \nabla_h \Psi(P) = E_P[h\phi_2]. \tag{3.2.4}$$

This of course implies that

$$E_P[(\phi_1 - \phi_2)h] = 0, \tag{3.2.5}$$

which is what we needed to show. □

Lemma 3.14. *Consider a statistical estimand $\Psi : \mathcal{M} \to \mathbb{R}$ and let $\mathcal{T}(P)$ denote the closure of the linear span of scores corresponding to the set of all possible differentiable paths through P in $\mathcal{M}$. The projection of any IF corresponding to a RAL estimator of Ψ onto the space $\mathcal{T}(P)$ is unique and equal to the EIF.*

Proof. Consider any IF ϕ and denote its projection onto $\mathcal{T}$ by $\phi^\dagger$, then

$$\phi = \phi^\dagger + h^\perp \tag{3.2.6}$$

for some $h^\perp \in \mathcal{T}^\perp$. By (3.2.6) it is clear that

$$E[h\phi] = E[h(\phi^\dagger + h^\perp)] = E[h\phi^\dagger], \tag{3.2.7}$$

which makes $\phi^\dagger$ an IF by Theorem 3.11. Hence, we know that there is an IF that is contained in $\mathcal{T}$.

To prove uniqueness, consider the projection $\tilde{\phi}^\dagger \in \mathcal{T}$ of another IF $\tilde{\phi}$ onto the tangent space $\mathcal{T}$, which by definition implies

$$\tilde{\phi} = \tilde{\phi}^\dagger + \tilde{h}^\perp \tag{3.2.8}$$

where $\tilde{h}^\perp \in \mathcal{T}^\perp$. As $\mathcal{T}$ is a linear subspace, we know that $\phi^\dagger - \tilde{\phi}^\dagger \in \mathcal{T}$. In addition, by Corollary 3.13 and the fact that we have just shown that $\phi^\dagger, \tilde{\phi}^\dagger$ are IFs it is clear that $\phi^\dagger - \tilde{\phi}^\dagger \in \mathcal{T}^\perp$ as well. However, as $\mathcal{T} \cap \mathcal{T}^\perp = \{0\}$ it must imply that $\phi^\dagger = \tilde{\phi}^\dagger$.

To show that the projection $\phi^\dagger$ is in fact the EIF, we consider the variance

$$\text{Var}[\phi] = \text{Var}\big[\phi^\dagger + h^\perp\big] = \text{Var}\big[\phi^\dagger\big] + \text{Var}\big[h^\perp\big] \geq \text{Var}\big[\phi^\dagger\big]. \tag{3.2.9}$$

Now we have shown that the IF obtained by projecting any IF onto $\mathcal{T}(P)$ must be the EIF as it satisfies (3.1.29). □

By Lemma 3.14, we know that we can find the EIF, as it is simply the projection of any other IF onto the tangent space. This is a powerful result as we are now able to construct a RAL estimator that is asymptotically efficient by building it around having IF equal to the EIF. One type of estimators are the plug-in estimators, which we will show is in fact RAL, under some regularity conditions, in the next section.

3.3 Plug-in estimators

In this project, we will mainly focus on a specific class of estimators called *plug-in estimators*. In short, a plug-in estimator is an estimator where an estimate $\hat{P}_n$ of the true distribution P_0 is plugged into the statistical estimand to obtain the parameter estimate $\hat{\Psi}_n = \Psi(\hat{P}_n)$. Plug-in estimators are advantageous since they respect the assumptions and constraints contained in the statistical model, and hence they are more robust to outliers, small samples and sparsity than non-plug-in estimators [11, ch. 1, 4].

In general we want asymptotic linear estimators, see Definition 3.3, and to obtain asymptotic linearity of a plug-in estimator, we need to show that it can be written as

$$\hat{\Psi}_n - \Psi(P_0) = \mathbb{P}_n \phi_{P_0} + o_{P_0}\left(n^{-1/2}\right). \tag{3.3.1}$$

Consider the true DGP P_0 and a fixed estimate of this distribution $\hat{P}_n$. Now, construct a path from $\hat{P}_n$ to P_0 defined by $\tilde{p}_\varepsilon = \varepsilon p_0 + (1-\varepsilon)\hat{p}_n$, as in Example 3.3 where $P = \hat{P}_n$ and $\tilde{P} = P_0$. Assume that Ψ is pathwise differentiable, then by using Definition 3.10 we can write

$$\Psi(P_0) - \hat{\Psi}_n = E_{\hat{P}_n}[\phi_{\hat{P}_n} h] + R(\hat{P}_n, P_0), \tag{3.3.2}$$

where R is a remainder term. By the change-of-variable formula and (3.1.20), we can write

$$\begin{aligned} E_{\hat{P}_n}[\phi_{\hat{P}_n} h] &= \int \phi_{\hat{P}_n}(o) h(o) d\hat{P}_n = \int \phi_{\hat{P}_n}(o) h(o) \hat{p}_n(o)\, do \\ &= \int \phi_{\hat{P}_n}(o)(p_0(o) - \hat{p}_n(o))\, do = \int \phi_{\hat{P}_n} dP_0 - \int \phi_{\hat{P}_n} d\hat{P}_n = P_0 \phi_{\hat{P}_n}, \end{aligned} \tag{3.3.3}$$

where $P_0\phi_{\hat{P}_n}$ denotes $E_{P_0}[\phi_{\hat{P}_n}]$, which is convenient later. Hence the estimation error is

$$\hat{\Psi}_n - \Psi(P_0) = -P_0\phi_{\hat{P}_n} - R(\hat{P}_n, P_0). \tag{3.3.4}$$

By adding zero in a smart way, we arrive at the following decomposition

$$\hat{\Psi}_n - \Psi(P_0) = \mathbb{P}_n \phi_{P_0} - \underbrace{\mathbb{P}_n \phi_{\hat{P}_n} + (\mathbb{P}_n - P_0)(\phi_{\hat{P}_n} - \phi_{P_0}) - R(\hat{P}_n, P_0)}_{x}. \tag{3.3.5}$$

To be asymptotically linear, x needs to be $o_{P_0}(n^{-1/2})$, which is satisfied if each of the terms in x are $o_{P_0}(n^{-1/2})$ individually.

The quantity $-\mathbb{P}_n\phi_{\hat{P}_n}$ will be referred to as the *plug-in bias*, which is often of concern when doing plug-in estimation. One way to eliminate the plug-in bias in the estimate is to just add the term again, shown in the following example. However it will be apparent in Chapter 4 that there are other ways of eliminating plug-in bias.

Example 3.7 One step estimator to eliminate plug-in bias
One way to construct an estimator that eliminates plug-in bias is to use the one step estimator

$$\hat{\Psi}_n^* = \hat{\Psi}_n + \mathbb{P}_n \phi_{\hat{P}_n}, \tag{3.3.6}$$

which is just a debiasing of the original plug-in estimate. Note that Ψ_n^* is not a plug-in

estimator. By (3.3.5) it is clear that

$$\hat{\Psi}_n^* - \Psi(P_0) = \mathbb{P}_n\phi_{P_0} + \underbrace{(\mathbb{P}_n - P_0)(\phi_{\hat{P}_n} - \phi_{P_0}) - R(\hat{P}_n, P_0)}_{x} \tag{3.3.7}$$

where we can conclude that the one step estimator is asymptotically linear with IF ϕ_{P_0} if we can show that $x = o_{P_0}(n^{-1/2})$.

The term

$$(\mathbb{P}_n - P_0)(\phi_{\hat{P}_n} - \phi_{P_0}). \tag{3.3.8}$$

in (3.3.5) is the *empirical process*. The convergence of this term requires a result from empirical process theory, which we will not go into detail with in this project. If we can show that $\phi_{\hat{P}_n}$ and ϕ_{P_0} fall into a P_0-*Donsker class* of functions and that $P_0(\phi_{\hat{P}_n} - \phi_{P_0})^2$ converges to zero in probability, we have that the empirical process is $o_{P_0}(n^{-1/2})$ by [22, Lemma 19.24]. We will not be going into the definition of Donsker classes in this project and from now on we will simply assume that this condition is satisfied. For more details see [23] and [24].

Lastly, for any given statistical estimation problem, we need to show that the remainder term

$$R(\hat{P}_n, P_0) = \hat{\Psi}_n - \Psi(P_0) + P_0\phi_{\hat{P}_n} \tag{3.3.9}$$

is $o_{P_0}(n^{-1/2})$, to obtain an asymptotic linear estimator. If p_0 is the density for the distribution P_0, we can sometimes factorise $p_0 = q_0 \cdot g_0$, and then let Q_0 denote the distribution with density q_0 and G_0 denote the distribution with density g_0. In the case when the remainder can be written as

$$R(\hat{P}_n, P_0) = \int (H_1(\hat{Q}_n) - H_1(Q_0))(H_2(\hat{G}_n) - H_2(G_0))f(\hat{P}_n, P_0)dP_0, \tag{3.3.10}$$

for functions H_1, H_2 and f, we say that the estimation problem admits a *double robustness structure*. The remainder converges to zero if either Q or G is consistently estimated. The rationale behind the name double robust lies in the fact that a consistent target parameter estimation depends on either one of them being consistent. In fact, if the remainder is as (3.3.10), it is sufficient that either $H_1(\hat{Q}_n)$ or $H_2(\hat{G}_n)$ is consistent. This will be illustrated in the following example.

Example 3.8 (Continuation of Example 3.5)
If we let P_0 denote the true DGP, then $\Psi(P_0)$ will denote the true value, ie. the target parameter. We will now consider the remainder term in (3.3.9), from the decomposition of the estimation error when estimating $\Psi(P_0)$ using a plug-in estimator. Since $\Psi(P_0) = \Psi_1(P_0) - \Psi_0(P_0)$ we will consider the remainder term for $\Psi_1(P_0)$ and the calculations are similar for $\Psi_0(P_0)$. Let $\hat{\bar{Q}}(A, W) = E_{\hat{P}_n}[Y \mid A, W]$, $\bar{Q}_0(A, W) = E_{P_0}[Y \mid A, W]$, $\hat{\bar{G}}(W) = E_{\hat{P}_n}[A \mid W]$ and $\bar{G}_0(W) = E_{P_0}[A \mid W]$. Using the EIF given in (3.1.32) it holds

that

$$\begin{aligned} R_1(\hat{P}_n, P_0) &= \Psi_1(\hat{P}_n) - \Psi_1(P_0) + P_0\phi^*_{1,\hat{P}_n} \\ &= E_{\hat{P}_n}\left[\hat{\bar{Q}}(1,W)\right] - E_{P_0}\left[\bar{Q}_0(1,W)\right] + \\ &\quad E_{P_0}\left[\frac{I(A=1)}{\hat{\bar{G}}(W)}(Y - \hat{\bar{Q}}(A,W)) + \hat{\bar{Q}}(1,W) - E_{\hat{P}_n}\left[\hat{\bar{Q}}(1,W)\right]\right] \end{aligned} \tag{3.3.11}$$

The very first and the very last expectation are equivalent but with opposite signs, hence they eliminate each other. It is possible to rewrite the term

$$\begin{aligned} E_{P_0}\left[\frac{I(A=1)}{\hat{\bar{G}}(W)}Y\right] &= E_{P_0}\left[E_{P_0}\left[\frac{I(A=1)}{\hat{\bar{G}}(W)}Y \;\middle|\; W\right]\right] \\ &= E_{P_0}\left[\frac{1}{\hat{\bar{G}}(W)}E_{P_0}[I(A=1)Y \mid W]\right] \\ &= E_{P_0}\left[\frac{1}{\hat{\bar{G}}(W)}E_{P_0}[A \mid W]E_{P_0}[Y \mid A=1,W]\right] \\ &= E_{P_0}\left[\frac{\bar{G}_0(W)}{\hat{\bar{G}}(W)}\bar{Q}_0(1,W)\right], \end{aligned} \tag{3.3.12}$$

where the third equation holds by Theorem A.8, and likewise

$$E_{P_0}\left[\frac{I(A=1)}{\hat{\bar{G}}(W)}\hat{\bar{Q}}(A,W)\right] = E_{P_0}\left[\frac{\bar{G}_0(W)}{\hat{\bar{G}}(W)}\hat{\bar{Q}}(1,W)\right]. \tag{3.3.13}$$

Hence the remainder (3.3.11) becomes

$$\begin{aligned} R_1(\hat{P}_n, P_0) &= E_{P_0}\Bigg[\frac{1}{\hat{\bar{G}}(W)}\Big(\bar{G}_0(W)\left(\bar{Q}_0(1,W) - \hat{\bar{Q}}(1,W)\right) \\ &\qquad + \hat{\bar{G}}(W)\left(\hat{\bar{Q}}(1,W) - \bar{Q}_0(1,W)\right)\Big)\Bigg] \\ &= E_{P_0}\left[\frac{1}{\hat{\bar{G}}(W)}(\hat{\bar{G}}(W) - \bar{G}_0(W))\left(\hat{\bar{Q}}(1,W) - \bar{Q}_0(1,W)\right)\right]. \end{aligned} \tag{3.3.14}$$

Note that this explicitly shows the doubly robustness property (3.3.10) of the estimation problem. That is, if either $\hat{\bar{G}}$ is consistently estimating $\bar{G}_0$ or $\hat{\bar{Q}}$ is consistently estimating $\bar{Q}_0$ then the remainder will be negligible. Since we are in an RCT setting, the treatment mechanism $\bar{G}_0$ is known, and hence the remainder term will be zero in this case by the above result.

4 Targeted Learning

In Chapter 2 we presented the statistical model, the statistical estimand, and arrived at a point where we obtained identifiability of a causal quantity through observable data under some assumptions. Turning back to the framework presented in Figure 1.1, we have now established the statistical estimation problem, and the purpose of this chapter is to construct a reasonable estimator with properties discussed in Chapter 3. For this purpose, we introduce Targeted learning and in particular Targeted Maximum Likelihood Estimation (TMLE), which has become increasingly popular since it was proposed by [20]. This is also for a good reason as, under reasonable regularity conditions, it yields an asymptotically efficient estimator of the statistical estimand. This chapter will describe the concept of TMLE and its properties, including an extension to the case of longitudinal data, and it is based on [11, ch. 4, 5] and [17].

In short, TMLE combines semiparametric efficiency theory, from Chapter 3, and machine learning in a two-step procedure, which can be summarised as follows.

(i) Obtain an initial estimate of the DGP or the relevant parts of this distribution using machine learning.

(ii) Update the initial estimate targeted towards making the optimal bias-variance trade-off for the target parameter.

This is solely meant to be a short overview of the general idea behind TMLE. To understand what is meant by relevant parts of the distribution, think of the estimand in Example 2.5 where we statically set the intervention nodes. Here we are only interested in the means under the conditional distribution of Y and the marginal distribution of W and not all of the distribution of $O = (W, A, Z, Y)$.

Machine learning are popular for some problems as they place few assumptions on the underlying distribution of the data and can accommodate a large number of covariates with complex relationships as opposed to parametric models. However, a concern about flexible data-adaptive models is that they do not have asymptotic properties for inference and are hence mainly used for prediction problems, however, prediction is not the aim when doing estimation. This is because machine learning fits have an optimal bias-variance trade-off for the estimation of the outcome, rather than an optimal bias-variance trade-off for the estimation of the parameters.

So how does one get the benefits from the flexible and assumption-cheap methods provided by machine learning but also the ability to obtain valid inference for parameters? This is where TMLE comes in. What makes TMLE different from the machine learning models is that after utilising machine learning for the initial estimate it takes an additional step that optimises the bias-variance trade-off for the target parameter instead of minimising prediction error. In the semiparametric setting with an infinite-dimensional parameter, maximum likelihood methods fail to work. However, TMLE proposes a way to still utilise maximisation of a likelihood when the model

space is semiparametric. This is done by iteratively parametrising a part of the model space using paths, and since paths have a one-dimensional parameter it enables utilisation of regular maximum likelihood estimation. As shown later, we obtain known asymptotic properties of bias and variance from this procedure enabling us to obtain valid inference from the TMLE estimate.

4.1 Targeted Maximum Likelihood Estimation

In this section, we will go through the steps of the TMLE algorithm, before extending the algorithm to handle longitudinal data. Suppose we have n observations, denoted $o_i = (w_i, a_i, z_i, y_i)$ for $i = 1, \dots, n$, of the stochastic vector $O = (W, A, Z, Y) \sim P_0$ with sample space $\mathcal{O}$. Consider a semiparametric statistical model $\mathcal{M}$, which is augmented with factual knowledge e.g. boundedness, independence or class of the variables, such that $P_0 \in \mathcal{M}$. Define a pathwise differentiable statistical estimand

$$\Psi : \mathcal{M} \to \mathbb{R}. \tag{4.1.1}$$

We will denote the true value of the parameter of interest by $\Psi(P_0)$.

The following algorithm constructs an estimator of the DGP, targets it towards the target parameter using the EIF, and obtain the estimate using a plug-in estimator. The target step serves to solve something called the efficient score equation and hence obtain the optimal bias-variance trade-off for $\Psi(P_0)$ [25].

Algorithm 4.1 Targeted Maximum Likelihood Estimation [11, ch. 5].

Require: A statistical model $\mathcal{M}$, a statistical estimand $\Psi : \mathcal{M} \to \mathbb{R}$ and n i.i.d. observations $o_1, \dots, o_n$ of a stochastic vector $O = (W, A, Z, Y) \sim P_0 \in \mathcal{M}$.

1: Obtain an initial estimate P_n^0 of P_0 with density p_n^0.
2: Set $i = 0$ and $\hat{\varepsilon} = 1$.
3: **while** $\hat{\varepsilon} \neq 0$ **do**
4: Compute the EIF $\phi_{P_n^i}$ corresponding to the statistical estimand at P_n^i.
5: Define a path $\{\tilde{P}_\varepsilon : \varepsilon \in [0, 1]\}$ with corresponding densities $\tilde{p}_\varepsilon = (1 - \varepsilon\phi_{P_n^i})p_n^i$.
6: Use maximum likelihood estimation along the path to update

$$\hat{\varepsilon} = \arg\max_{\varepsilon} \frac{1}{n} \sum_{i=1}^{n} \log(\tilde{p}_\varepsilon(O_i)), \tag{4.1.2}$$

where we define $\log(0) := -\infty$.
7: Update $P_n^{i+1} = \tilde{P}_\varepsilon|_{\varepsilon=\hat{\varepsilon}}$ and denote its density by p_n^{i+1}.
8: Update $i = i + 1$.
9: The estimate obtained in the final iteration, denoted by P_n^*, is the TMLE of the true DGP P_0. This is used to obtain the TMLE of $\Psi(P_0)$ by simply plugging the final distribution estimate into the statistical estimand:

$$\hat{\Psi}_{\text{TMLE}} = \Psi(P_n^*). \tag{4.1.3}$$

This algorithm requires some explanation, which are listed in the following:

- In each iteration in the algorithm, we consider a one-dimensional maximum likelihood problem in the parameter ε, which means that in each step we will increase the empirical

log-likelihood, relative to the value in the previous step. The main purpose of this algorithm is to adjust the initial estimate in a reasonable way such that the plug-in bias is eliminated.

- The initial estimate in step 1 can be obtained by using any model, for example a parametric regression or a flexible data-adaptive machine learning method. Computation of the EIF in step 4 is performed using the theory described in Section 3.2.
- Note that as the estimate changes in step 7 so does the path in step 5 in the next iteration. Since the new path starts in the estimate obtained in the previous iteration, we are guaranteed that the log-likelihood value does not decrease when maximising along this new path in step 6.
- In step 5, the likelihood is continuous as a function of ε and as ε takes values on a closed interval, it is also bounded and hence it attains a maximum for $\varepsilon \in [0, 1]$. Since the likelihood is bounded from above, and as just explained, the value of the likelihood never decreases even though it changes in each iteration, we will reach the maximum at some point implying that $\hat{\varepsilon} = 0$.

This algorithm, where the likelihood is maximised to obtain the estimate, is a special case of a broader class of estimators, called targeted minimum loss-based estimators, which is also abbreviated TMLE. This class of estimators works under minimisation of a loss function, which satisfies that the minimum expected value is obtained in the truth. Hence in the presented algorithm, the loss function will be defined as the negative log-likelihood, which attains its minimum expected value in the true DGP.

One can imagine that the statistical estimand only depends on some part Q_0 of P_0. In this case we would typically specify the loss function such that it also only depends on that part of the distribution. In this case we could replace P with Q in the Algorithm 4.1.

There might be some concerns regarding the validity of the steps in Algorithm 4.1, which are listed and will be discussed in the following:

- One concern is that the functions defined in step 5 might not be valid densities. That is, there is a chance that there exists an $\varepsilon \in [0, 1]$ and an $o \in \mathcal{O}$ such that $\tilde{p}_\varepsilon(o) < 0$. However, if for an arbitrary $P \in \mathcal{M}$ it holds that $\phi^*_P : \mathcal{O} \to \mathbb{R}$ is continuous and $\mathcal{O}$ is a compact set, then there exists some $M \in \mathbb{R}$ such that $\phi^*_P(o) \geq -M$. By Proposition 3.7, we know that this immediately implies that $\tilde{p}_\varepsilon$ is a well-defined density for all $\varepsilon \in [0, 1]$ if $M = 1$. Following the proof of Proposition 3.7 we know that $\tilde{p}_\varepsilon$ is a well-defined density for all $\varepsilon \in [0, \min\{1, 1/M\}]$ when $\phi^*_P(o) \geq -M$ for all $o \in \mathcal{O}$. That is, we can adjust the interval for ε such that we always consider valid densities and hence valid distributions. In practice, ensuring a lower bound for $\phi^*_{P^i_n}$ is in general not a concern as it is based on a finite amount of observed data, which must make up a compact set.
- In the case of a semi-parametric model there might also be a concern regarding whether or not the densities from step 5 correspond to distributions that reside in the model $\mathcal{M}$. To avoid this problem we will assume that there always exists an $\mathcal{E} \in \mathbb{R}$ such that $\varepsilon < \mathcal{E}$ implies $\tilde{P}_\varepsilon \in \mathcal{M}$ [12].
- Note that we constructed the path using the EIF, and not the score as in Proposition 3.7. However, by Lemma 3.14 the EIF is in the tangent space making it a score, and hence defines a valid direction for a path. By using this path in the maximising step of the algorithm, we are moving in the direction in which we obtain the largest possible change in the estimate towards the truth $\Psi(P_0)$. Any other direction may give small adjustments in the estimate, but larger changes in parts of the distribution that do not contribute to the estimate, which are

not of interest. This is derived from a discussion on the hardest submodel or locally least favorable model, which we will not dive deeper into in this project. For more context see [12, 4.4].

In the extreme case where both of the first two concerns are applicable, we will maximise over the values of ε that ensure we have a well-defined density and then stay within $\mathcal{M}$.

It remains to show that Algorithm 4.1 gives a useful estimate of $\Psi(P_0)$, which will be elucidated in the next section. It turns out that the TMLE algorithm is a recipe for constructing an estimator with IF equal to the EIF and hence a recipe for constructing an asymptotically efficient estimator.

4.2 Properties

From (4.1.3) it is clear that TMLE is a plug-in estimator. To obtain valid and sensible inference, it is important to have some control over how this estimator behaves asymptotically. In the following, we will show that TMLE eliminates plug-in bias, which makes it an asymptotically linear estimator under some regularity conditions as discussed in Section 3.3.

As explained in Subsection 3.1.1, asymptotic linearity is a very attractive property for an estimator to have, and hence we are looking to see if the estimator obtained by using the Algorithm 4.1 possesses this property. Recall that the decomposition of the estimation error (3.3.5) holds for any plug-in estimator. By the arguments made in Section 3.3, we need to consider the size of the plug-in bias for this case. The following result shows that even though TMLE is a plug-in estimator it eliminates plug-in bias.

Proposition 4.1. *The TMLE P_n^* of P_0 eliminates plug-in bias*

$$\mathbb{P}_n \phi_{P_n^*} = \frac{1}{n} \sum_{i=1}^{n} \phi_{P_n^*}(o_i) = 0, \tag{4.2.1}$$

where $\phi_{P_n^*}$ denotes the EIF of the statistical estimand at P_n^*.

Proof. Let p_n^* denote the density of P_n^*. Consider the path $\{\tilde{P}_\varepsilon^* : \varepsilon \in [0,1]\}$ with densities $\tilde{p}_\varepsilon^* = (1 + \varepsilon \phi_{P_n^*}) p_n^*$, then

$$\log(\tilde{p}_\varepsilon^*) = \log((1 + \varepsilon \phi_{P_n^*}) p_n^*) \tag{4.2.2}$$

and hence

$$\frac{d}{d\varepsilon} \log(\tilde{p}_\varepsilon^*) = \frac{1}{1 + \varepsilon \phi_{P_n^*}} \phi_{P_n^*} \quad \Longrightarrow \quad \frac{d}{d\varepsilon} \log(\tilde{p}_\varepsilon^*)|_{\varepsilon=0} = \phi_{P_n^*}. \tag{4.2.3}$$

When the maximiser of the log-likelihood is 0, it holds that

$$\mathbb{P}_n \frac{d}{d\varepsilon} \log(\tilde{p}_\varepsilon^*)|_{\varepsilon=0} = 0, \tag{4.2.4}$$

where $\mathbb{P}_n$ is the empirical mean, as introduced in Section 3.1. Hence, it becomes clear that the TMLE P_n^* eliminates plug-in bias

$$\mathbb{P}_n \phi_{P_n^*} = 0 \tag{4.2.5}$$

by combining (4.2.3) and (4.2.4). □

Based on this result and the discussion in Section 3.3, it holds that TMLE is an asymptotically linear estimator that provides efficient and consistent estimates for a doubly robust estimation problem where at least one of the components in the second order remainder term is estimated consistently under regularity conditions, including that the EIF falls in a P_0-Donsker class. In addition, it can be shown that the TMLE algorithm also yields a regular estimator, which we will not go into in this project [11].

This section proves that it is possible to construct an efficient estimator directly from knowing the EIF. Hence, we can use the theory from Chapter 3 to identify the EIF and Algorithm 4.1 to construct an efficient estimator from the estimation problem described in Chapter 2.

5 Data and methods

To explore the problem described in Chapter 1 regarding estimation of the treatment effect when participants initiate rescue medication during clinical trials, we have been provided with data from Novo Nordisk A/S. The data are from the phase 3a programme PIONEER, which consisted of many different trials, each with a different aspect in mind. We will only consider the PIONEER 1 trial, which was introduced in Section 1.3 and will be explained in greater detail in the following section. In addition, we will describe methods for estimating the hypothetical estimand in context of the PIONEER 1 trial. First, some rather simple methods will be proposed and then we describe the method that is standard of practice for estimating the hypothetical estimand. Lastly, we will describe how one can use targeted learning to estimate the hypothetical estimand. The results from these methods will be compared in Chapter 6 and 7.

5.1 PIONEER 1

In this section we will describe relevant aspects, setup, inclusion criteria and guidance on rescue medication in context of the PIONEER 1 trial. PIONEER 1 is a multicentre randomised 26-week placebo-controlled monotherapy trial designed to assess the clinical effects of oral semaglutide at the three intended dose levels (3, 7 and 14 mg) vs. placebo in subjects with type 2-diabetes (T2D) who at trial entry were being treated with diet and exercise only [6]. For a brief introduction to what oral semaglutide is, how it treats diabetes and an explanation of most important biomarkers measured in the study, we refer the reader to Section 1.3 or [8].

In the PIONEER 1 trial, the participants show up at the site for assessment of their measurements at four planned visits between the baseline visit at week 0 and the final visit at week 26. The visit structure in the trial, along with which variables are collected at that visit and the corresponding threshold for receiving rescue medication, is summarised in Table 5.1.

Table 5.1: *Summarising the visit structure for PIONEER 1, along with the variables collected at each visit and the threshold for rescue medication if it exists.*

Visit #	Weeks	Variables	Rescue threshold
Visit 0	Week 0	W_0, A_0, Z_0	
Visit 1	Week 4	W_1, A_1, Z_1	
Visit 2	Week 8	W_2, A_2, Z_2	FPG > 13.3
Visit 3	Week 14	W_3, A_3, Z_3	FPG > 11.1
Visit 4	Week 20	W_4, A_4, Z_4	FPG > 11.1
Visit 5	Week 26	Y	

In general, the Ws and the Y denote continuous measurements of HbA_{1c}, bounded between 0

and 100%, at the corresponding visit. In the PIONEER 1 study, the baseline value of HbA_{1c} is prespecified to be adjusted for and hence it is the first entry in the vector of baseline covariates W_0. The distribution of the baseline HbA_{1c} levels can be found in Figure B.3. The only other covariate prespecified for use in the primary analysis, is the covariate `REGION`, which is a categorical variable that indicates the continent of residence and its distribution can be seen in Figure B.2. The As are binary variables capturing adherence to the randomised treatment, except A_0 which is a binary indicator of randomised treatment that is equal to 1 if assigned to active treatment. For example, $A_2 = 0$ if the participant discontinued treatment before visit 3, 14 weeks after randomisation. We assume, in line with the protocol, that once a participant has discontinued the randomised treatment, they can not go back. That is, if $A_k = 0$ it implies that $A_j = 0$ for all $j > k$. Finally, the Zs are binary variables that tell whether or not the participants receive rescue medication between that visit and the next visit. Just as for the As, if $Z_k = 1$ it implies that $Z_j = 1$ for all $j > k$. A summary of these variables and the notation that will be used throughout the rest of this project is given in Table 5.2.

Table 5.2: *Overview of the different variables in the analysis dataset, their name in the code and their notation.*

Description	**Variable name**	**Notation**
HbA_{1c} at baseline	`HBA1CBL`	$W_{0,1}$
Region of residence	`REGION`	$W_{0,2}$
Randomised treatment	`A0`	A_0
Rescue medication before visit 1	`Z0`	Z_0
Change from baseline in HbA_{1c} at visit 1	`HBA1CV1`	W_1
Treatment discontinuation before visit 2	`A1`	A_1
Rescue medication before visit 2	`Z1`	Z_1
Change from baseline in HbA_{1c} at visit 2	`HBA1CV2`	W_2
Treatment discontinuation before visit 3	`A2`	A_2
Rescue medication before visit 3	`Z2`	Z_2
Change from baseline in HbA_{1c} at visit 3	`HBA1CV3`	W_3
Treatment discontinuation before visit 4	`A3`	A_3
Rescue medication before visit 4	`Z3`	Z_3
Change from baseline in HbA_{1c} at visit 4	`HBA1CV4`	W_4
Treatment discontinuation before visit 5	`A4`	A_4
Rescue medication before visit 5	`Z4`	Z_4
Change from baseline in HbA_{1c} at visit 5	`HBA1CV5`	Y

The variables in the data and their mutual dependencies can be visualised in a DAG, presented in Figure 5.1. To illustrate the dependencies of one of the variables some arrows are marked with red in Figure 5.1. The node which we will focus on is the node W_2, the measurement of HbA_{1c} at visit 2, 8 weeks after randomisation. The coloured arrows illustrate all the node's parents. That is, W_2 is dependent on W_0, A_0, Z_0, W_1, A_1, and Z_1, all the history prior to visit 2. As another example, the parents of Y is all of the variables except for Y itself.

The DAG, presented in Figure 5.1 is rather complex as it barely imposes any assumptions on the relationships between the variables, except for the time ordering. Since the variables we are working with are measurements of HbA_{1c}, the average level of blood sugar over the past three months, it may be reasonable to assume that a measurement is primarily affected by the most recent measurement and not previous ones. Likewise, discontinuation of randomised treatment and rescue medication intake are most likely not affected by past events if we know the most recently measured

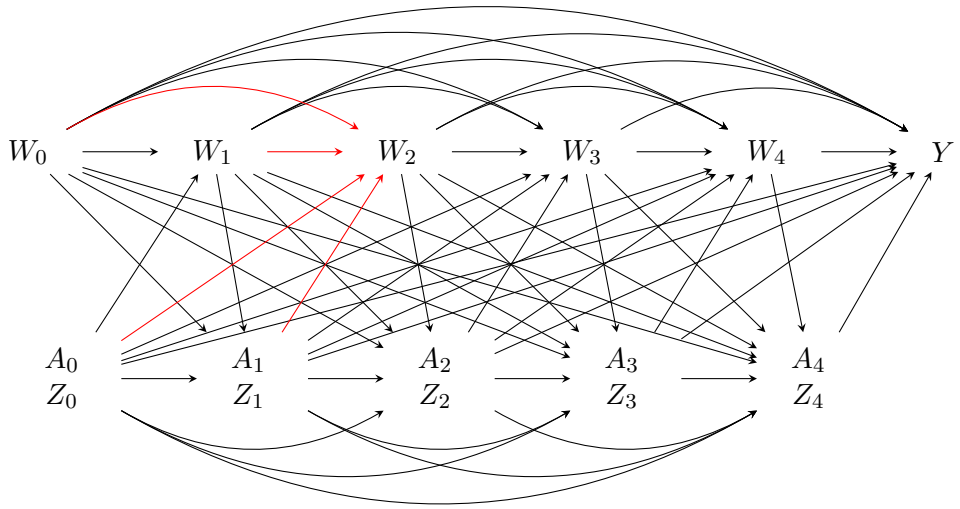


Figure 5.1: *A joint DAG illustrating the dependencies in the data from the PIONEER 1 trial. Red arrows mark the route from the parents of the node W_2.*

variables. Hence we will also propose the DAG presented in Figure 5.2. This DAG makes a Markov type of assumption implying conditional independencies between many pairs of variables. Making more assumptions in this way might introduce bias as we might remove dependencies that actually exist, however, it will most likely lower the variance making it part of the bias-variance trade-off. In addition, the assumptions made by removing these arrows from the DAG are reasonable as it just implies that the most recent events have a direct effect and the events prior to that have an indirect effect through the most recent events.

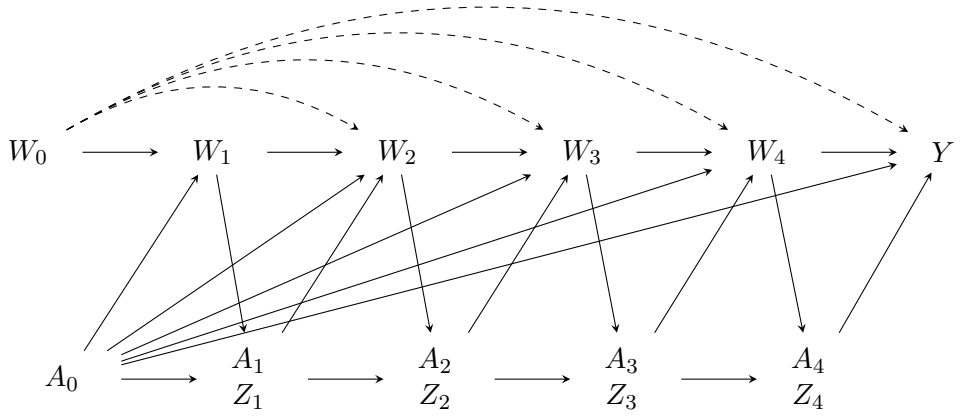


Figure 5.2: *A joint DAG illustrating the dependencies in the data from the PIONEER 1 trial after making some assumptions that simplify the dependency structure. Dotted lines coming from W_0 indicate that only the region, $W_{0,2}$, affects the variable that the arrow is pointing to.*

At each visit, the physician at the site will collect a blood sample from the participant measuring the relevant information. Specifically, the physician will among other things receive measurements of the HbA_{1c} and the FPG. The HbA_{1c} value is contained in the W corresponding to that specific visit and the mean value of these across participants for each visit is displayed in Figure 5.3. The value of

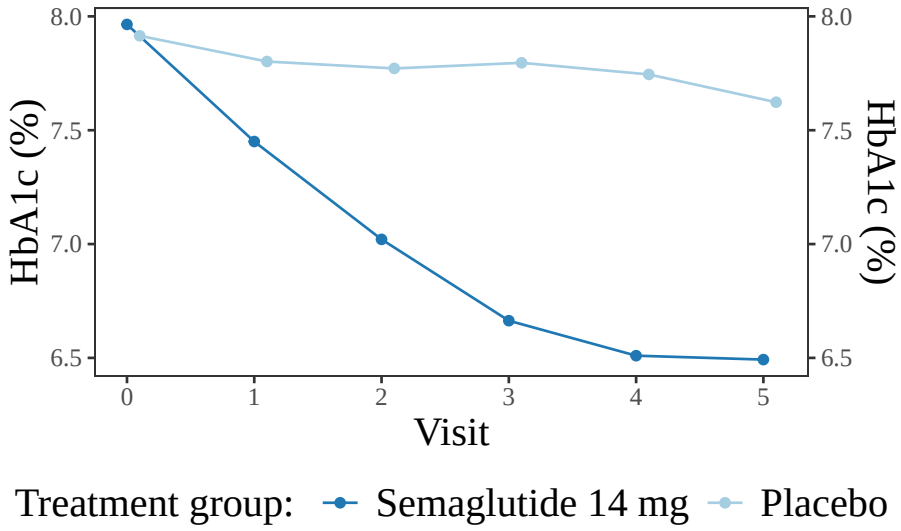


Figure 5.3: *Mean plot for each treatment group along with the standard error.*

FPG is used to determine the need for rescue medication, hence the value of Z based on thresholds presented in Table 5.1. These thresholds align with an example of prespecified thresholds provided by FDA through [26], who also stress the need for a period of around 6 weeks in the beginning of a trial, to allow time for the treatment to work. Hence rescue medication is generally not allowed in this period, but it is important to consider later on in the trial since participants with persistent and unacceptable hyperglycaemia should be offered treatment intensification [6]. From now on, in the analysis of PIONEER 1 and in the simulation study, the variable Z_0 will be omitted. This is because participants are not allowed to enter the trial with rescue medication, and since the first time glycaemic control is assessed, after receiving the randomised treatment, is at visit 1.

The biomarkers HbA_{1c} and FPG measure the same phenomenon, namely the level of blood sugar. Due to the nature of HbA_{1c}, which captures average blood glucose during the last two to three months, it makes sense to use this biomarker to evaluate the efficacy of the drug. By the same argument, it makes less sense to use this variable to determine the need for rescue medication, since by construction it is less sensitive to recent changes. However, FPG is a measurement that captures the level of blood sugar at the moment that it is measured. By nature, this variable has a lot more variability, and hence it is better to capture recent changes. By the same arguments, this variable is not a suitable biomarker for estimating the efficacy.

Naturally, there are some missing observations in the dataset from the PIONEER 1 trial. This project will not be addressing methods for missing data and we will use multiple imputation to complete the data prior to fitting models. In order to do this, we use the `mice` package in R.

Prior to the analyses we displayed histograms of baseline variables grouped by treatment group in Figure B.2 and Figure B.3. In addition, on page 93-95 in [6] two tables summarising descriptive statistics on baseline characteristics and demographics can be found. Both the figures and the tables show equal distributions of baseline variables across the two treatment groups as expected.

5.2 The Estimation problem

Section 5.1 gave a description of the experiment that created the data that we will be analysing in this project and the potential causal relationships among each variable involved. This leads us to

consider the clinical question of interest, which we can then formulate as a causal quantity and lastly a statistical estimation problem following the workflow illustrated in Figure 1.1.

Clinical question of interest

To assess the efficacy of oral semaglutide compared to placebo, the clinical question of interest was

What is the treatment effect of oral semaglutide 14 mg compared to placebo in patients with T2D assuming that all subjects remained on trial product and did not use rescue medication? [6]

In order to more formally formulate this question, we describe it in terms of a clinical estimand, which has five attributes that were described in Section 1.2.

Clinical estimand

As explained, we are looking into different ways of estimating the hypothetical estimand, which is the treatment effect where intercurrent events are handled using the hypothetical strategy, described in Section 1.2. In [6] they state

The hypothetical estimand evaluated the treatment effect assuming that all subjects remained on trial product and did not initiate any rescue medication. Thus, this estimand estimates the efficacy for a scenario where the drug is taken as intended and where rescue medication is not initiated.

The case described in the quote above is exactly what the clinical question of interest is pointing toward and hence we formulate this as a clinical estimand:

- **Target population:** Participants aged ≥ 18 years with T2D for at least 30 days, who at trial entry were being treated with diet and exercise only and had HbA_{1c} levels between 7% and 9.5%. For thorough inclusion and exclusion criteria see Table B.1.
- **Treatment:** Oral semaglutide 14 mg compared to placebo.
- **Outcome variable:** Change in HbA_{1c} from baseline to week 26 per the discussion in Section 5.1.
- **Population-level summary:** Mean difference, in change in HbA_{1c} from baseline, between treatment groups.
- **Strategies for handling ICEs:** Hypothetical strategy for treatment discontinuation and rescue medication use.

Using these five attributes, we have clearly specified exactly what we would like to know. Now, we are ready to move on to the mathematical formulation of the quantity of interest. This will be done within the Rubin causal model framework in terms of potential outcomes as introduced in Subsection 2.2.1.

Causal question of interest

Before defining the Rubin causal model for this problem, we will introduce some notation. Let $\mathbb{V}_k$ denote the vector of potential outcomes of the variable V_k at visit k for any variable $V \in \{W, Y\}$ in every possible treatment scenario with respect to time ordering and dependencies. In addition, let $V_{j:k}$ for $j \leq k$ denote the collection of variables $(V_j, \ldots, V_k)$ for any variable $V \in \{W, A, Z\}$. As an example, $\mathbb{W}_1 = (W_1(A_0 = 0), W_1(A_0 = 1))$ as the only treatment that can affect W_1 is A_0, which only has two possible values. However, this becomes complicated really fast if we consider the dependencies from Figure 5.1 as W_3 is affected by A_0, A_1, A_2, Z_1 and Z_2, which can attain 2

different values each, resulting in $2^5 = 32$ potential outcomes. We will not bother to write these out and hence use the notation $\mathbb{W}_3$ for this collection of potential outcomes instead. In addition, there are $2^9 = 512$ different potential outcomes of Y under the DAG in Figure 5.1.

In our case, only two of the potential outcomes for Y are of interest as for each of the treatments we only want to consider the case where participants remain on the randomised treatment $A_{1:4} = \mathbb{1}_4 = (1, 1, 1, 1)$ while not taking rescue medication $Z_{1:4} = \mathbb{0}_4 = (0, 0, 0, 0)$ at any time during the study. This will be more clear when we introduce the causal estimand of interest in the following.

Rubin causal model

Data	The data in the causal world consists of observations from the stochastic vector $$O^* = (W_0, A_0, \mathbb{W}_1, A_1, Z_1, \ldots, \mathbb{W}_4, A_4, Z_4, \mathbb{Y}),$$ using the notation introduced above.
Model	A causal model $\mathcal{M}^*$, which is a collection of potential distributions of O^*, designed to comply with the natural constraints implied by data, eg. boundedness, independence and class of the variables.
Estimand	The causal estimand of interest is $$\Psi^*(P) = E_P[Y(1, \mathbb{1}_4, \mathbb{0}_4) - Y(0, \mathbb{1}_4, \mathbb{0}_4)]$$ for $P \in \mathcal{M}$, where $Y(a_0, \mathbb{1}_4, \mathbb{0}_4)$ denotes the potential outcome at the final visit when the randomised treatment A_0 is set to $a_0 \in \{0, 1\}$ and $A_{1:4} = \mathbb{1}_4$ and $Z_{1:4} = \mathbb{0}_4$ corresponding to adhering to the randomised treatment and not initiating rescue medication throughout the study.

Identifiability

Identifiability of the causal estimand by a statistical one is ensured by generalising the identifying assumptions of no unmeasured confounders, positivity and consistency in Assumption 2.11 to the longitudinal case. The observable data structure in the PIONEER 1 trial is

$$O = (W_0, A_0, W_1, A_1, Z_1, \ldots, W_4, A_4, Z_4, Y) \in \mathcal{O}. \tag{5.2.1}$$

Assumption 5.1 (Longitudinal identifying assumptions)**.**

(i) **Sequential randomisation:**
- $[Y(1, \mathbb{1}_4, \mathbb{0}_4), Y(0, \mathbb{1}_4, \mathbb{0}_4)] \perp\!\!\!\perp A_k \mid W_{0:k}, A_{0:(k-1)}, Z_{1:(k-1)}$
- $[Y(1, \mathbb{1}_4, \mathbb{0}_4), Y(0, \mathbb{1}_4, \mathbb{0}_4)] \perp\!\!\!\perp Z_k \mid W_{0:k}, A_{0:k}, Z_{1:(k-1)}$

for all $k = 0, \ldots, K-1$.

(ii) **Sequential positivity:**

$$\sup_{O \in \mathcal{O}} \prod_{k=0}^{4} \frac{I(A_k = 1)}{g_{A_k}} \frac{I(Z_k = 0)}{g_{Z_k}} < \infty \tag{5.2.2}$$

where g_{A_k} and g_{Z_k} denote the conditional densities $p(A_k \mid \text{Pa}(A_k))$ and $p(Z_k \mid \text{Pa}(Z_k))$ respectively. Here we define $0/0 := 1$.

(iii) **Sequential consistency:** The potential outcome $Y(a_0, \mathbb{1}_4, \mathbb{0}_4)$ is observed for the subjects that followed the treatment regime $A_0 = a_0, A_{1:4} = \mathbb{1}_4, Z_{1:4} = \mathbb{0}_4$.

In the following we will go through each of the assumptions in Assumption 5.1 and discuss the validity in the context of the PIONEER 1 trial. The sequential randomisation assumption is satisfied as long as there are no unmeasured confounders in the outcome-treatment relationship, where A_0, A_k and Z_k for $k = 1, \ldots, 4$ are considered treatments and Y is considered the outcome. It has been discussed in a variety of papers on diabetes studies, see [2] and [27], that repeated measurements of both HbA_{1c} and FPG collected before the outcome Y are confounders as they affect both the occurrence of an ICE and the value of the outcome. We will not use FPG as a covariate, hence it will be considered unmeasured in our analyses making it an unmeasured confounder. However, when discussing the validity of the assumption of no unmeasured confounders, [2] states "*the extent to which it is violated may be mitigated by the high correlation between FPG and HbA*$_{1c}$". As both FPG and HbA_{1c} are blood glucose measurements they are highly correlated, which makes it less of a concern not to include FPG in our models even though it is believed to be a confounder.

It is clear that the sequential positivity assumption, Assumption 5.1 (ii), is a bit different and less intuitive than what we have seen before, but it is simply an extension. In the context of the PIONEER 1 trial, sequential positivity corresponds to having a positive probability of observing the outcome in the case of no discontinuation of treatment and no rescue medication under either randomised treatment for every participant. Hence, we consider the product of ratios between the interventions, determined by the indicators $I(A_k = 1)$ and $I(Z_k = 0)$, and the true conditional densities of A_k and Z_k. The sequential positivity assumption requires that every intervention $A_k = 1$ and $Z_k = 0$, for $k = 1, \ldots, 4$, has positive probability of occurring in the true distribution, just as in Assumption 2.11 (ii). We know that there is a positive probability of either treatment scenario by randomisation. Moving on to treatment discontinuation, we need a positive probability of not discontinuing the randomised treatment at any time no matter the covariate values. The concern is that there might be participants that experience adverse events or insufficient therapeutic effect, which makes them want to discontinue treatment. When looking into different variables in the dataset and comparing extreme values that might affect discontinuation of trial product, we saw no patterns or trends that led to a violation of this assumption.

The procedure outlined in Section 5.1, regarding initiation of rescue medication based on a threshold, highlights that there might be an issue with the assumption of having a positive probability of not receiving rescue medication at any visit. According to the protocol it seems that initiation of rescue medication is a deterministic decision based on covariate history. We have investigated this assumption in data, which shows some deviations from this deterministic decision guiding framework. In the PIONEER 1 trial participants were not allowed to initiate rescue medication before visit 2. However there were some (< 5) participants that initiated rescue medication before visit 2. In later visits, where there were thresholds for initiation of rescue medication as indicated in Table 5.1, at least 50% of the participants who exceeded these thresholds did not end up receiving rescue medication. For visit 2, 3 and 4, the proportions of participants that actually received rescue medication among those exceeding the rescue threshold are 38.1%, 45.7% and 21.6% respectively. This may be because they did not end up exceeding the threshold when a confirmatory FPG measurement was taken, but we have not investigated this part of the problem. The conclusion is, that the threshold rule in the protocol did not end up creating a deterministic rescue medication assignment after exceeding certain thresholds of FPG when based only on scheduled visits, as we might have feared. Hence we are no longer concerned about the validity of the assumption of having positive probability of not receiving rescue medication for every possible covariate history. Of course we need a positive probability of the combination of no rescue medication and continuing on trial product throughout the 26 weeks for both treatment arms. However, it is not believed that these two ICEs affect each other and hence it is sufficient to consider them separately.

The sequential consistency assumption, Assumption 5.1 (iii), states that the participant's realised

outcome should be equal to the potential outcome corresponding to the treatment scenario they have experienced. For example, there should not be a variation of treatment dose within one treatment arm. As doses are measured exactly in the PIONEER 1 trial, we have no reason to doubt the validity of this assumption.

In conclusion, the assumptions stated in Assumption 5.1 unreasonable. If there is a concern of any assumptions being valid, one could conduct a sensitivity analysis in addition to the original analysis, however, this is out of scope for this project.

Statistical problem

We are able to move on to describe the statistical estimation problem as illustrated in Figure 1.1 and under the identifying assumptions stated in Assumption 5.1, it holds that the statistical estimand given by

$$\begin{aligned}\Psi(P) = E[E[\cdots E[Y \mid W_{0:4}, A_0 = 1, A_{1:4} = \mathbb{1}_4, Z_{1:4} = \mathbb{0}_4]\cdots \mid W_0, A_0 = 1]] \\ - E[E[\cdots E[Y \mid W_{0:4}, A_0 = 0, A_{1:4} = \mathbb{1}_4, Z_{1:4} = \mathbb{0}_4]\cdots \mid W_0, A_0 = 0]]\end{aligned} \tag{5.2.3}$$

identifies the causal estimand given above. We will not be proving this result, however, the proof of an analogue result in the simpler case where there are 2 visits after baseline is given in Appendix A.4.

The statistical estimation problem

Data	n i.i.d. observations $o_1, \ldots, o_n$ of the stochastic vector $$O = (W_0, A_0, W_1, A_1, Z_1, \ldots, W_4, A_4, Z_4, Y)$$ using the notation introduced earlier.
Model	A statistical model $\mathcal{M}$, which is a collection of distributions designed to comply with the natural constraints implied by data, eg. boundedness, independence and class of the variables.
Estimand	The statistical estimand of interest $\Psi : \mathcal{M} \to \mathbb{R}$ is given in (5.2.3).

The last steps presented in the workflow of Figure 1.1 are to define a statistical estimator, obtain a statistical estimate and interpret the result.

5.2.1 Data for analysis of the hypothetical estimand

As introduced in Section 5.2, we are interested in the hypothetical estimand, that is, we are interested in the case where participants do not initiate rescue medication nor discontinue the randomised treatment throughout the study. In the redacted clinical study report [6], they define different observation periods used for efficacy endpoints. The two observation periods that we will consider in this project are

- Data from randomised participants while on treatment and without rescue medication
- Data from all randomised participants in trial

The first approach is to solely consider pre-ICE data. That is, discarding data observed after occurrence of any of these ICEs, creating a monotone missingness pattern. The intuitive idea to this approach is to eliminate the effect that ICEs have on the outcome by disregarding observations that are affected by them. In the context of the PIONEER 1 study, observations after treatment discontinuation or initialisation of rescue medication are discarded and the corresponding variables

A_k and Z_k are referred to as *censoring variables*, as they together are indicators of whether or not an observation is missing. The second approach is to utilise all the collected data, model the effects of the ICEs on the outcome and then calculate the estimate of the treatment effect in the hypothetical absence of ICEs.

In the following sections, we will present methods that use data from either of these observation periods and compare them in a simulation study and in a case study, see Chapters 6 and 7.

5.3 Parametric estimation methods

First we will propose likelihood based methods, which assume parametric dependencies and only use data from participants while on treatment and without rescue medication. When using likelihood based methods on data with missingness in the dependent variable, we work under the assumption that data is missing at random (MAR) [28]. This corresponds to non-informative censoring, that is, the censoring could not have been foreseen, had we known the outcomes after censoring. When basing these models on data observed prior to any ICE, we answer the question "*What would the treatment effect be, had patients taken the randomised treatment and not received rescue medication and behaved like other patients who did not take rescue medication and adhered to protocol?*". In words, the MAR assumption states that the behaviour of participants that experience an ICE is similar to the behaviour of participants that do not experience any ICEs and hence we can generalise the results drawn from pre-ICE data to all participants.

To state the MAR assumption mathematically in the context of basing an analysis on pre-ICE data, we start by introducing some notation. In PIONEER 1 there were 5 visits after baseline, hence we define the joint censoring variable

$$R_k = \begin{cases} 0, & \text{if either } A_k = 0 \text{ or } Z_k = 1 \\ 1, & \text{otherwise} \end{cases} \tag{5.3.1}$$

for $k = 1, \ldots, 4$, which are indicators of whether or not data after the k'th visit are included in the analysis. By the monotone missingness pattern that is created by discarding post-ICE data, $R = (R_1, \ldots, R_4)$ can only take the form

$$r_j = (\mathbb{1}_j, \mathbb{0}_{4-j}) \tag{5.3.2}$$

for $j = 0, \ldots, 4$. Then $R = r_j$ for one subject would imply that we solely include data observed up to visit j and consider everything observed thereafter as missing. That is $R = r_4$ denotes the case where the participant does not discontinue randomised treatment nor initialise rescue medication at any point during the study and hence there is no missingness in (W_2, W_3, W_4, Y).

By [29], the MAR assumption states

$$P(R = r_j \mid A_0, W_{0:(j+1)}, W_{(j+2):4}, Y) = P(R = r_j \mid A_0, W_{0:(j+1)}) \tag{5.3.3}$$

for $j = 0, \ldots, 3$. This is equivalent to the conditional independence

$$\{R = r_j\} \perp\!\!\!\perp (W_{(j+2):4}, Y) \mid (A_0, W_{0:(j+1)}). \tag{5.3.4}$$

Part of this assumption is somewhat familiar to us as (5.3.4) is similar to the sequential randomisation assumption stated in Assumption 5.1 [30]. As discussed earlier, this assumption simply entails that we do not have any unobserved confounders in the relationship between the outcome and the time varying treatment variables A_k and Z_k. However, the MAR assumption additionally requires that there are no unobserved confounders in the relationship between the repeated measurements of the endpoint up to the final visit and the time varying treatment variables A_k and Z_k.

5.3.1 Simple methods modelling only the outcome

If one just considers the outcome Y at end of treatment and hence ignore all the repeated measurements $W_{1:4}$, the outcome will be censored if one of the $A_k = 0$ or $Z_k = 1$, and due to the monotonicity of these variables it is equivalent with $R_4 = 0$. In this case, where we just ignore the information between the baseline visit and the final outcome measurement, the MAR assumption simplifies to the conditional independence

$$\{R_4 = 0\} \perp\!\!\!\perp Y \mid (A_0, W_0). \tag{5.3.5}$$

The implication of this assumptions is, that we receive no information from Y as to whether or not a participant experience any ICE. However, we know that the prescription of rescue medication is determined by the biomarker FPG, which is highly correlated with the HbA_{1c} measurement. The repeated measurements of HbA_{1c} are therefore considered confounders in the R_4-Y relationship, which is also depicted in both Figure 5.1 and 5.2. Hence, we do not believe this assumption, but use it as a starting point for estimation of the treatment effect.

In this subsection we will consider some simple likelihood based estimation methods that solely consider the final outcome as the dependent variable and only consider data that are not affected by any ICE. In this case, participants that discontinue at the first visit and participants who discontinue just before the last visit contribute with the same amount of information. These methods rely on the MAR assumption as discussed above.

Empirical mean

The first simple approach is to take the empirical mean of the final HbA_{1c} measurement across each of the treatment groups, that is

$$\bar{Y}_0 = \frac{1}{n_0} \sum_{i=1}^{n} I(A_{0,i} = 0) Y_i, \quad \bar{Y}_1 = \frac{1}{n_1} \sum_{i=1}^{n} I(A_{0,i} = 1) Y_i, \tag{5.3.6}$$

where n_0 and n_1 are the number of participants that do not discontinue treatment or receive rescue medication in the placebo and treatment group respectively. The estimate of the treatment effect using this approach is $\bar{Y}_1 - \bar{Y}_0$. Under the assumption that the outcome is normally distributed this is a likelihood based approach, since it is equivalent to fitting a linear model where the dependent variable is Y and the only independent variable is A_0.

Linear model

Another simple approach, using the same strategy for handling ICEs, is to fit a linear model accounting for covariates. We will adjust for the baseline value of HbA_{1c} and region of residence inspired by the covariates that were adjusted for in the original analysis. We specify the linear model

$$Y = \gamma_0 + \gamma_1 W_{0,1} + \gamma_2 W_{0,2} + \gamma_3 A_0 + \varepsilon, \tag{5.3.7}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma^2 > 0$ and one should recall the notation introduced in Table 5.2. In this case, the treatment effect is γ_3.

5.3.2 Mixed models for repeated measures

In addition to the simple methods presented above, we will present the method that was chosen for the hypothetical estimand when originally planning the PIONEER 1 study, which is a Mixed

Model for Repeated Measures (MMRM). The MMRM models a continuous variable that has been measured repeatedly at discrete time points and can include both fixed and random effects. Hence we can utilise the repeated measures from data on treatment and without rescue medication, $\Upsilon = (W_1, W_2, W_3, W_4, Y)$. In this case, the participants contribute with different amounts of information if they are censored at different time points, contrary to the case where we only consider the outcome. The advantage of using an MMRM is that it allows us to account for the correlation between the repeated measurements of HbA_{1c} and hopefully gain information about the trajectory of the censored participants. This method is standard practice for estimating the hypothetical estimand as it provides unbiased results under the assumption that missing data is MAR and the model is correctly specified [31]. We aim to replicate the analysis from the publication [32]. In the case where there is no missingness, which in this case corresponds to no ICEs, MMRM is equivalent to a general linear model.

As our aim is to replicate the original analysis, we will just state the model. The supplementary material for [32] formulated the MMRM model in the following way: "*The independent effects included in the model were treatment and region as categorical fixed effects and baseline value as a covariate, all nested within visit. An unstructured covariance matrix for endpoint measurements within the same patient was employed*". This means that the fixed effects consisted of HbA_{1c} measured at baseline, region, visit, randomised treatment A_0 and the interactions between visit and all the other variables. Let Υ_k denote the k'th entry in Υ. Using the notation presented in Table 5.2, the model described above can be formulated in the following way

$$\Upsilon_k = \alpha_0 + \beta_{0,k} + W_{0,1}\beta_{1,k} + W_{0,2}\beta_{2,k} + A_0\beta_{3,k} + \varepsilon_k \tag{5.3.8}$$

for each post-baseline visit $k = 1, \ldots, 5$, where we assume that $\sum_{k=1}^{5} \beta_{0,k} = 0$. The error term ε_k denotes the k'th entry in $\varepsilon \sim \mathcal{N}_5(\mathbb{0}_5, \Sigma)$. As (5.3.8) contains no random effects, it is simply a general linear model. The model makes the assumption that the change in HbA_{1c} measured at each visit throughout the duration of the trial can be described by a finite number of parameters. We have earlier discussed that this assumption might not be very realistic and in practice, a sensitivity analysis is often conducted to ensure that the findings are not completely reliant on this and other perhaps inaccurate assumptions.

As region is a factor with five levels, we assume that $W_{0,2}$ is a four dimensional row vector of indicators that specify the region that the participant resides in, where a full row of zeros indicate that the participant resides in the reference group, which is important to avoid overparameterisation. The parameters $\beta_{2,k} \in \mathbb{R}^4$ are four dimensional parameter vectors for $k = 1, \ldots, 5$ specifying the effect of each region and its interaction with the visits.

The covariance matrix $\Sigma \in \mathbb{R}^{5\times5}$ is assumed to be unstructured meaning that we have 15 variance parameters to estimate. When using an unstructured covariance matrix, we make no assumptions about the structure of correlational effects. A drawback of using an unstructured covariance matrix is that we need to estimate a large number of parameters compared to other covariance structures that make more restrictive assumptions. However, this is not of concern in large samples. In total we reach 50 parameters to estimate when fitting this model.

The MMRM (5.3.8) assumes that the residuals are multivariate normally distributed. This is again a restrictive assumption, but we can assess a normal Quantile-Quantile (QQ) plot for the residuals as a rather simple tool to see if the residuals at each separate visit look normally distributed. Figure B.1 displays the normal QQ plots for the residuals at each visit after baseline. It is clear that they all have rather heavy tails, but they fit fairly well towards the middle. Judging from these plots there is not enough evidence to reject the assumption of normally distributed data.

As the aim is to model the data as if it had not been affected by treatment discontinuation or

rescue medication intake with an MMRM, we consider the data structure (W_0, A_0, Υ). That is, we disregard the censoring variables A_k and Z_k, which we can do as we have made the assumption that these variables give no further information about the value of the potential outcomes of a participant. By MAR (5.3.4), the statistical estimand (5.2.3) equals

$$\begin{aligned}&E[\cdots E[Y \mid W_{0:4}, A_0 = 1] \mid W_{0:3}, A_0 = 1] \mid W_{0:2}, A_0 = 1] \mid W_{0:1}, A_0 = 1] \mid W_0, A_0 = 1]] \\ &- E[\cdots E[Y \mid W_{0:4}, A_0 = 0] \mid W_{0:3}, A_0 = 0] \mid W_{0:2}, A_0 = 0] \mid W_{0:1}, A_0 = 0] \mid W_0, A_0 = 0]] \\ &= E[E[Y \mid W_0, A_0 = 1] - E[Y \mid W_0, A_0 = 0]] = \beta_{3,5}.\end{aligned}$$

The first equality is a consequence of Theorem A.7 and the second is per (5.3.8). The treatment effect that we will record for the MMRM model will be the estimate of the above quantity.

As we have mentioned earlier in this project, there are quite a few drawbacks when using parametric models like MMRM, as correct specification of the model is an important assumption for it to attain desirable properties. In a guideline specifically addressing the hypothetical strategy for the use of rescue medication in the context of diabetes mellitus [33], EMA stated that "*standard imputation methods or modelling targeting a hypothetical estimand strategy may not be appropriate if based on subjects that do not require rescue medication or if based on the missing-data-assumption since these subjects differ from those who require rescue medication*". This motivates looking into other approaches for estimating the hypothetical estimand, which do not rely on the MAR assumption. In addition, we might wonder if there is a method where it is possible to utilise all the observed data, instead of ignoring the parts that are influenced by treatment discontinuation and initialisation of rescue medication.

5.4 Longitudinal TMLE

As opposed to the parametric estimation methods we will introduce a method based on TMLE that is semiparametric and utilises all the observed data. First we need to extend TMLE, from Section 4.1, to handle longitudinal data.

Recall the statistical estimand of interest presented in Section 5.2,

$$\begin{aligned}\Psi(P) = E[&E[\cdots E[Y \mid W_{0:4}, A_0 = 1, A_{1:4} = \mathbb{1}_4, Z_{1:4} = \mathbb{0}_4] \cdots \mid W_0, A_0 = 1]] \\ &- E[E[\cdots E[Y \mid W_{0:4}, A_0 = 0, A_{1:4} = \mathbb{1}_4, Z_{1:4} = \mathbb{0}_4] \cdots \mid W_0, A_0 = 0]].\end{aligned} \tag{5.4.1}$$

We start by introducing some notation that will come in handy when extending the concept of TMLE to this setting. Recall that we have used $\bar{Q}$ to denote the conditional mean of the outcome Y. Since we are working iteratively we will subscript the Qs to indicate which mean we are considering. For $a_0 \in \{0, 1\}$, let

$$\bar{Q}_Y(a_0, W_{0:4}) \coloneqq E[Y \mid W_{0:4}, A_0 = a_0, A_{1:4} = \mathbb{1}_4, Z_{1:4} = \mathbb{0}_4] \tag{5.4.2}$$
$$\bar{Q}_{W_4}(a_0, W_{0:3}) \coloneqq E[\bar{Q}_Y(a_0, W_{0:4}) \mid W_{0:3}, A_0 = a_0, A_{0:3} = \mathbb{1}_3, Z_{1:3} = \mathbb{0}_3] \tag{5.4.3}$$
$$\bar{Q}_{W_3}(a_0, W_{0:2}) \coloneqq E[\bar{Q}_{W_4}(a_0, W_{0:3}) \mid W_{0:2}, A_0 = a_0, A_{1:2} = \mathbb{1}_2, Z_{1:2} = \mathbb{0}_2] \tag{5.4.4}$$
$$\bar{Q}_{W_2}(a_0, W_{0:1}) \coloneqq E[\bar{Q}_{W_3}(a_0, W_{0:2}) \mid W_{0:1}, A_0 = a_0, A_1 = 1, Z_1 = 0] \tag{5.4.5}$$
$$\bar{Q}_{W_1}(a_0, W_0) \coloneqq E[\bar{Q}_{W_2}(a_0, W_{0:1}) \mid W_0, A_0 = a_0]. \tag{5.4.6}$$

Longitudinal TMLE (LTMLE) is a process that iteratively uses the TMLE algorithm for each of the equations (5.4.2)-(5.4.6). When iterating, it moves backwards in time and starts by estimating (5.4.2) using the TMLE algorithm to find an updated estimator $\hat{\bar{Q}}^*_Y(a_0, W_{0:4})$ of $\bar{Q}_Y(a_0, W_{0:4})$. Then the obtained estimator is used as the outcome in the next conditional expectation (5.4.3),

which we use TMLE to find an estimator for. This estimator is then used as the outcome in the next conditional expectation and so on until we get an updated estimator $\hat{\bar{Q}}^*_{W_1}(a_0, W_0)$ of $\bar{Q}_{W_1}(a_0, W_0)$. Then the LTMLE of $\Psi(P_0)$ is defined by

$$\hat{\Psi}_{\text{LTMLE}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\bar{Q}}^*_{W_1}(1, w_{0,i}) - \hat{\bar{Q}}^*_{W_1}(0, w_{0,i}) \tag{5.4.7}$$

where $w_{0,1}, \ldots, w_{0,n}$ denote i.i.d. observations of W_0. This procedure is described in the following algorithm.

Algorithm 5.1 LTMLE for the intervention specific mean.

Require: A statistical model $\mathcal{M}$ containing distributions with densities that are bounded from above, the statistical estimand $\Psi : \mathcal{M} \to \mathbb{R}$ in (5.4.1) and n i.i.d. observations $o_1, \ldots, o_n$ of the stochastic vector $O = (W_0, A_0, W_1, A_1, Z_1, \ldots, W_4, A_4, Z_4, Y) \sim P_0 \in \mathcal{M}$.

1: Use Algorithm 4.1 to estimate $\bar{Q}_Y(a_0, W_{1:4})$ and denote the TMLE of $\bar{Q}_Y(a_0, W_{1:4})$ by $\hat{\bar{Q}}^*_{W_5}$
2: **for** $k = 4$ **to** $k = 1$ **do**
3: Obtain the TMLE $\hat{\bar{Q}}^*_{W_k}$ of $\bar{Q}_{W_k}(a_0, W_{1:k})$ by Algorithm 4.1, using $\hat{\bar{Q}}^*_{W_{k+1}}$ as the outcome.
4: Save the final estimator of $\Psi(P_0)$ given by (5.4.7).

Algorithm 5.1 uses Algorithm 4.1 iteratively, in which the EIF is calculated. The EIF in the k'th iteration is given by

$$\phi^{W_k} = \frac{I((A_0, A_{1:(k-1)}, Z_{1:(k-1)}) = (a_0, \mathbb{1}_{k-1}, \mathbb{0}_{k-1}))}{g_{A_0}\prod_{i=1}^{k-1} g_{A_i} g_{Z_i}}(\bar{Q}_{W_{k+1}} - \bar{Q}_{W_k}) \tag{5.4.8}$$

for $k = 1, \ldots, 4$ where $g_{A_k} = p(A_k \mid \text{Pa}(A_k))$ and $g_{Z_k} = p(Z_k \mid \text{Pa}(Z_k))$ and

$$\phi^{Y} = \frac{I((A_0, A_{1:4}, Z_{1:4}) = (a_0, \mathbb{1}_4, \mathbb{0}_4))}{g_{A_0}\prod_{i=1}^{4} g_{A_i} g_{Z_i}}(Y - \bar{Q}_Y) \tag{5.4.9}$$

for the first step in Algorithm 5.1 and a proof of this statement can be found in [34, app. A]. As TMLE is used repeatedly we know by Proposition 4.1 that plug-in bias is eliminated for each iteration considered as a separate problem. However, by [34, thm. 2] it holds that plug-in bias is eliminated for estimation of $\Psi(P_0)$ using LTMLE if either the estimates of (5.4.2)-(5.4.6) or $g = g_{A_0}\prod_{i=1}^{4} g_{A_i} g_{Z_i}$ are equal to the truth.

Algorithm 5.1 is a very specific version of LTMLE for estimating an intervention specific mean, but there are more general versions of this algorithm that are suitable for other statistical estimands. In this project we will solely consider intervention specific means, for which this algorithm is sufficient.

An advantage of using LTMLE for estimating the hypothetical estimand is that we are able to design the statistical model ourselves. That is, we can restrict our model space to reflect factual knowledge instead of making unrealistic assumptions, eg. parametric dependencies. As mentioned, we can use DAGs to illustrate the potential dependencies among variables in the data and hence specify the dependencies that are allowed in possible DGPs contained in the statistical model. When making less assumptions in the DAG, that is, having more potential dependencies, we will most likely end up with more uncertainty. Hence it is often of interest to remove some of these dependencies from the DAG as we did in Figure 5.2. However, it is a trade-off situation as these assumptions might not be satisfied, which will cause bias.

This concludes the presentation of the different methods that we will use for estimating the hypothetical estimand in this project. The aim of the next couple of chapters is to compare the results that each of these yield in a simulation study and a case study.

6 Simulation study

Causal estimands require unobserved counterfactuals and seek properties of the true data generating process, which remains unknown. When we seek to validate or compare models, this is quite an obstacle, since we do not know the truth and hence do not know what to compare the models to. Instead of using observed data where the DGP is unknown, we will conduct a simulation study where we choose the true DGP beforehand. Then we are able to compare models based on their performance on this simulated dataset, where we know the actual truth that our models are trying to estimate. This can be perceived as an imaginary world with all the information that we could wish for.

To investigate the difference between the performance of the models proposed in 5 for estimating the hypothetical estimand in a variety of scenarios, we conducted a simulation study. The scenarios we are interested in looking into are cases where the proportion of participants receiving rescue medication varies. This section will go through how the data are simulated and the conclusions that we are able to draw from this simulation study. All analyses were carried out using R version 4.4.1.

6.1 Distribution of simulated data

For the purpose of the simulation study, we developed a function in R to simulate longitudinal RCT data. To mimic data from the PIONEER 1 data, we simulated 1000 two-armed RCTs with 400 participants, where the continuous outcome was measured at baseline and at 5 visits after baseline.

First we simulated four baseline covariates, denoted $W_0 = (W_{0,1}, W_{0,2})$. The first baseline covariate will be interpreted as the baseline HbA_{1c} value and is chosen to be normally distributed with mean 7.94 and variance 0.49 inspired by the data from PIONEER 1. The second baseline covariate is a categorical variable indicating region of residence among five possible regions chosen at random such that two of the regions are chosen with a larger probability than the rest. In addition to these we simulated some additional baseline variables $U_0 = (U_{0,1}, U_{0,2})$ that will be considered unmeasured covariates in the analyses and will hence not be accounted for. The covariates $U_{0,1}$ and $U_{0,2}$ are chosen to be independent standard normal distributed variables. The treatment variable A_0 is evenly distributed. This concludes the variables that were generated for each participant at the baseline visit. Note that we did not allow participants to initiate rescue medication before visit 1. This choice was made as we do not see this in the data from PIONEER 1 and this extreme case is often not relevant in diabetes studies.

Visit 1

Moving on to the first visit, we start by generating the HbA_{1c} value by the following equation

$$\begin{aligned} W_1 = W_{0,1} &+ 0.1(U_{0,1} + U_{0,2}) + \theta_1 A_0 + 0.5 \cdot I(W_{0,2} = 3) \\ &+ 0.05(W^{\text{st}}_{0,1} U_{0,1} + W^{\text{st}}_{0,1} U_{0,2} + U_{0,1} U_{0,2}) - 0.05({W^{\text{st}}_{0,1}}^2 + U^2_{0,1} + U^2_{0,2}) + t_4, \end{aligned} \tag{6.1.1}$$

where $W_{0,1}^{\text{st}}$ denotes the standardised value of $W_{0,1}$ and t_4 denotes a t distribution with 4 degrees of freedom. The effect of treatment, which is determined by the parameter $\theta_1 = -0.5$ and number of visits K, is chosen to be negative such that the active treatment reduces the HbA_{1c} value, which is what we would expect in a diabetes trial. Based on the consultation in visit 1, participants may discontinue the randomised treatment and/or initiate rescue medication. We generate the variables

$$A_1 \sim \text{Bernoulli}(0.98) \tag{6.1.2}$$

$$Z_1 \sim \text{Bernoulli}(p_1), \quad p_1 = \text{expit}(\gamma_{0,1} + \gamma_{1,1}A_0 + \gamma_{2,1}W_1). \tag{6.1.3}$$

This implies that the probability of discontinuing treatment between the first and second visit ($A_1 = 0$) is the same across all participants and that the probability of receiving rescue medication varies depending on randomised treatment assignment, A_0, and the previously measured HbA_{1c} value, W_1. The coefficients $(\gamma_{0,1}, \gamma_{1,1}, \gamma_{2,1})$ were found by fitting a logistic regression on the data from the PIONEER 1 trial in order to mimic the way participants receive rescue medication between visits 1 and 2 in the study. The HbA_{1c} values in the data were standardised prior to fitting the regression and likewise W_1 was standardised prior to making the predictions $\text{expit}(\gamma_{0,1} + \gamma_{1,1}A_0 + \gamma_{2,1}W_1)$ that form the probabilities p_1 for each participant. Table 6.1 shows the estimated coefficients for the logistic regression and hence the values of $(\gamma_{0,1}, \gamma_{1,1}, \gamma_{2,1})$.

Remaining visits

Now that we have explained thoroughly how the variables for the baseline visit and visit 1 are generated we will use generalised notation to explain how variables observed in visits 2 through 4 are generated, as this is a repetitive process. For visit $k = 2, 3, 4$ we first generate the HbA_{1c} value observed for that visit using the following formula:

$$W_k = W_{k-1} + \frac{\theta_1(6-k)}{5}A_0A_{k-1} + \theta_2 Z_{k-1}I(Z_{1:k-2} = \mathbb{0}_{k-2}) + t_4. \tag{6.1.4}$$

The intuition of the formula (6.1.4) for the HbA_{1c} value at visit k is, that it is based on the previous HbA_{1c} value with additional effects depending on treatment assignment, treatment adherence and rescue medication intake. The indicator variable is added to make sure that the effect of taking rescue medication is only added once.

Then we generate the adherence and rescue medication intake variables. For the adherence variable we assume that when a participant does not adhere, it is not possible to get back on randomised treatment, that is $A_k = 0 \Rightarrow A_{k+1} = 0$. When they adhere, the variable A_k is generated in the same way as (6.1.2). For the rescue medication intake variable the rule is that you are only allowed rescue medication once, hence if a participant has not received anything yet, $Z_{k-1} = 0$, the Z_k is generated by

$$Z_k \sim \text{Bernoulli}(p_k), \quad p_k = \text{expit}(\gamma_{0,k} + \gamma_{1,k}A_0 + \gamma_{2,k}W_k) \tag{6.1.5}$$

and otherwise if $Z_{k-1} = 1$ then $Z_k = 1$. Remark, that a participant who has been assigned to active treatment, but discontinues treatment at some point, will progress as if they had been assigned to placebo after they discontinue. The coefficient $\theta_2 = -0.7$ meaning that participants who initiate rescue medication receive an immediate effect of reducing their HbA_{1c} value by 0.7. Lastly Y, the HbA_{1c} value at the end of treatment, is generated by

$$Y = W_4 + \frac{\theta_1}{5}A_0A_4 + \theta_2 Z_4 I(Z_3 = 0) + t_4. \tag{6.1.6}$$

6.1.1 Dependencies

As mentioned, the coefficients $(\gamma_{0,k}, \gamma_{1,k}, \gamma_{2,k})$ in the formulas for the probability of receiving rescue medication at the k'th visit for each participant are found using a logistic regression fit on the real data. The coefficients obtained from these fits are displayed in Table 6.1. The intercepts $\gamma_{0,k}$ are close to -10 for all $k = 1, \ldots, 4$ indicating that with a HbA_{1c} value of 0 at visit k, a participant in the placebo group would have an extremely small probability for receiving rescue medication at all visits. Of course this scenario is completely hypothetical as a HbA_{1c} level of 0 is unfeasible. The coefficient $\gamma_{1,k}$ is negative for all k, meaning that being assigned to the active treatment decreases the probability of getting rescue medication compared to placebo. This indicates that there will be created an imbalance of rescue medication use between the treatment groups. It is clear that receiving active treatment has a much larger effect on the probability of getting rescue medication for the first two visits ($k = 1, 2$) than for the last two visits ($k = 3, 4$). Lastly, the coefficient $\gamma_{2,k}$ is positive for all k implying that a larger HbA_{1c} value at the previous visit gives a higher probability of receiving rescue medication. This is expected as larger HbA_{1c} values indicate high values of blood glucose, which can lowered by rescue medication.

Table 6.1: *Coefficient estimates after fitting a logisitic regression, on data from the PIONEER 1 trial, with formula* $Z_k = \gamma_{0,k} + \gamma_{1,k}A_0 + \gamma_{2,k}W_k$.

k	$\gamma_{0,k}$	$\gamma_{1,k}$	$\gamma_{2,k}$
1	-10.2500	-17.0876	0.6278
2	-11.805	-17.652	1.062
3	-10.705	-2.016	1.046
4	-8.3217	-1.7680	0.8137

Using the above formulas for generating observations for each of the 400 participants in each dataset, we know the true causal distribution P^* of $O^* = (W_0, A_0, \mathbb{W}_1, A_1, Z_1, \ldots, \mathbb{W}_4, A_4, Z_4, \mathbb{Y})$. This means that we can calculate the true value of the causal estimand described in Section 5.2. Considering the scenario where participants completely adhere and never initiate rescue medication, the true treatment effect is given by

$$\begin{aligned}\Psi^*(P^*) &= E_{P^*}[Y(1, \mathbb{1}_4, \mathbb{0}_4) - Y(0, \mathbb{1}_4, \mathbb{0}_4)] \\ &= \sum_{k=1}^{K} \frac{K-k+1}{K}\theta_1 = \sum_{k=1}^{5} \frac{6-k}{5}(-0.5) = -1.5\end{aligned} \tag{6.1.7}$$

and is identified by the statistical estimand (5.2.3), which we will try to estimate using the methods described in Chapter 5 in the next section.

As we are in complete control of the underlying DGP, we have complete knowledge of the SCM, which is simply the collection of equations presented in the above and hence we can also draw a correct DAG for this case. Figure 6.1 presents the true relationships between the variables for the simulated data. Here it is important to note that A_k for $k = 2, \ldots, 4$ is only affected by A_{k-1} if $A_{k-1} = 0$, which then sets $A_k = 0$ deterministically. Otherwise A_k is completely randomly generated.

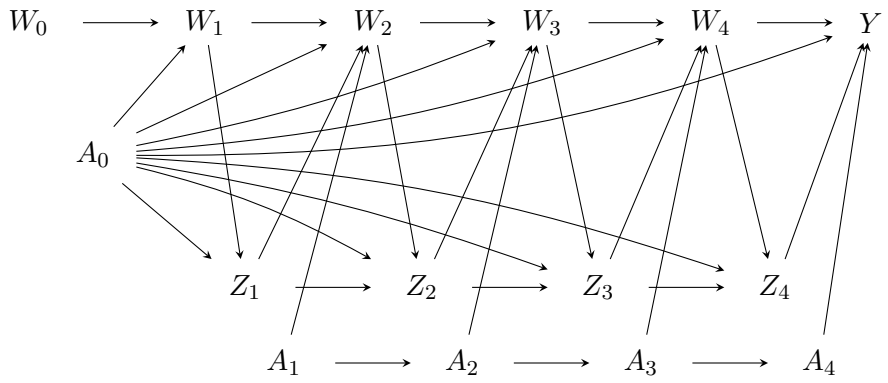


Figure 6.1: *DAG illustrating dependencies between variables in the simulated data.*

The DAG presented in Figure 6.1 is noticeably different than the DAGs in Figure 5.1 and 5.2. There are much fewer dependencies between the variables in the simulation study than what Figure 5.1 assumes, which ensures that a model containing all possible distributions with these dependencies will contain the true DGP of the simulated data. The DAG in Figure 5.2 does not present any dependencies between the randomised treatment A_0 and the ICEs A_k and Z_k for $k = 2, \ldots, 4$ as opposed to the simulated data where A_0 directly affects Z_k for $k = 1, \ldots, 4$. However, this was solely done in the simulated data to create the imbalance in the proportion of participants receiving rescue medication across treatment groups. In addition, it is not always the case that the assumptions made on dependencies between variables is completely correct and it will also be of interest to see how models that assume different dependencies end up performing, when these assumptions are not satisfied.

To get an understanding of how the different methods perform on average, we generate 1000 datasets consisting of 400 participants using the above formulas. From now on, we will refer to this collection of datasets as *scenario B*. On average across these 1000 datasets, 22.4% of the participants end up receiving rescue medication at some point during the trial. This is more than what was observed in the PIONEER 1 study, however, this will work in our favour as we are interested in how well the different methods presented in Chapter 5 handle data that is affected by intercurrent events.

6.1.2 Additional datasets

In addition to considering scenario B, we also generated 1000 datasets that was not affected by rescue medication at all, which will be referred to as *scenario A*. To generate these datasets, the same process as described in Section 6.1 was used while deterministically setting $Z_k = 0$ for all $k = 1, \ldots, 4$. This scenario is of interest because we would like to compare the performance of the methods presented in Chapter 5 when data is affected very little by ICEs. The only ICE in this case is treatment discontinuation, which is balanced among treatment groups by construction. Hence this dataset will represent the scenario where rescue medication was not permitted without any other changes to the DGP.

As we are interested in methods for estimating the hypothetical estimand, it is natural to question the performance of these in the case where more than 22.4% of the participants receive rescue medication. Hence we have also generated 1000 datasets where, on average, 38% of participants receive rescue medication, which will be referred to as *scenario C*. Again, we modified the amount of rescue medication use and used the same process for generating data as described in Section 6.1.

In Table 6.2 we have displayed the percentage of participants that receive rescue medication at some point for each of the three collections of 1000 datasets. The difference in the percentage of participants that take rescue medication across the two treatment groups is remarkable for scenario B and C. It is clear that the effect of treatment on the probability of receiving rescue medication that was presented in Table 6.1 shows in the simulated data just as expected. It should be noted that the true value of the hypothetical estimand is -1.5 in all of the generated datasets and it is only the percentage of participants that receive rescue medication that differs among each of the three collections of 1000 datasets.

Table 6.2: *The mean, minimum and maximum percent of participants (written as mean*$[\min, \max]$*) that receive rescue medication at some point throughout the study, taken over three different rescue scenarios, named A, B and C, each containing 1000 simulated datasets of 400 participants.*

Scenario	Treatment		Placebo		Total	
A	0%	$[0, 0]$	0%	$[0, 0]$	0%	$[0, 0]$
B	7.376%	$[2, 14.50]$	37.35%	$[25, 50]$	22.36%	$[14.75, 31]$
C	15.83%	$[9, 23.5]$	60.25%	$[50, 70.5]$	38.04%	$[30.75, 44.75]$

Figure 6.2 displays the average percent of participants that have not experienced an ICE at each given visit. That is, the percent of participants that are still taking the randomised treatment, $A_k = 1$, and have not started rescue medication, $Z_k = 0$, at each visit $k = 1, \ldots, 5$. As explained, data was simulated such that no participant discontinued treatment or initialised rescue medication before visit 1, making this percentage 100% for all three rescue scenarios. In scenario A, where no rescue medication was allowed, we only see a decline due to the participants that discontinue treatment, as this is the only ICE. Hence we end up with an outcome measurement Y, that was not affected by an ICE, from approximately 92% of participants on average across datasets from scenario A. In scenario B and C we see that there are approximately 72% and 57% of participants having an outcome measurement that is not affected by an ICE, marked by the dots at visit 5.

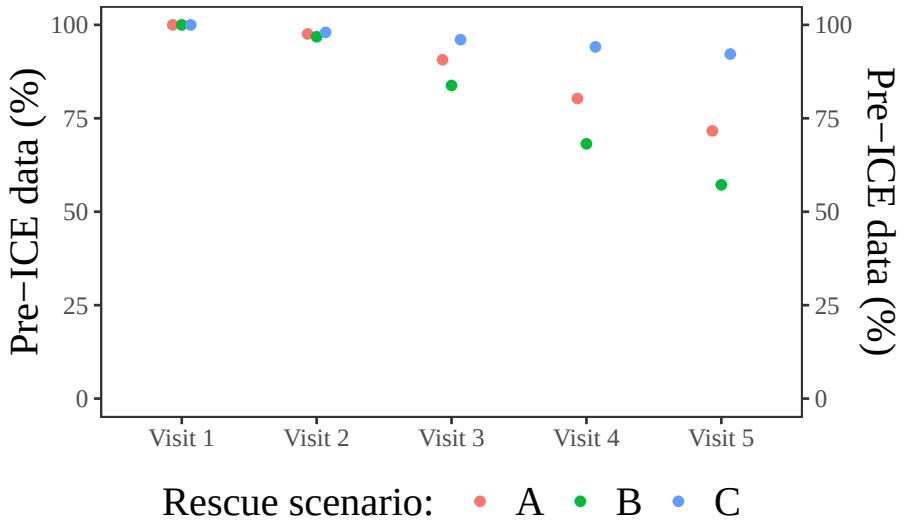


Figure 6.2: *Visualising the amount of data which are not affected by the occurrence of an ICE for each visit. Dots are slightly jittered in the direction of the x-axis for better visualisation.*

To get an idea of the mean trajectories of participants generated in each of the scenarios A, B and C we present Figure 6.3. These plots display the means across all 1000 datasets containing 400

participants, that is, they are means across 400,000 measurements for each of the visits. It is clear across all three plots that the treatment arm obtains a lower HbA_{1c} than the placebo arm. One thing to notice is that the placebo arms are rather different across the three plots, that is, the placebo arms for the different simulated datasets are different. This is due to the varying amount of rescue medication. Of course, there is also a varying amount of rescue medication in the treatment arm, however, as displayed in Table 6.2, the difference is much mere extreme in the placebo arm, which also makes up the noticeable difference in the plots.

In the first plot displayed in Figure 6.3 there is not much change in the placebo arm. In the second plot, there is a downhill slope after visit 2 indicating that placebo patients in these simulated trials actually lower their HbA_{1c} levels throughout the study. However, by the way this data has been generated, we know that this is due to the amount of participants in the placebo arm that end up receiving rescue medication, which by definition is some anti-diabetic drug that naturally will affect the HbA_{1c} measurements. Hence if we do not account for this difference, we will most likely underestimate the true treatment difference of the experimental treatment. In the last plot displayed in Figure 6.3, we see a more extreme case, where the proportion of participants receiving rescue medication is amplified. This has a clear effect on the mean trajectory of the placebo patients. Hence, in a case where a large proportion of participants end up receiving rescue medication, there is an even larger risk of underestimating the true treatment difference if the models used in the analysis do not account for the use of rescue medication.

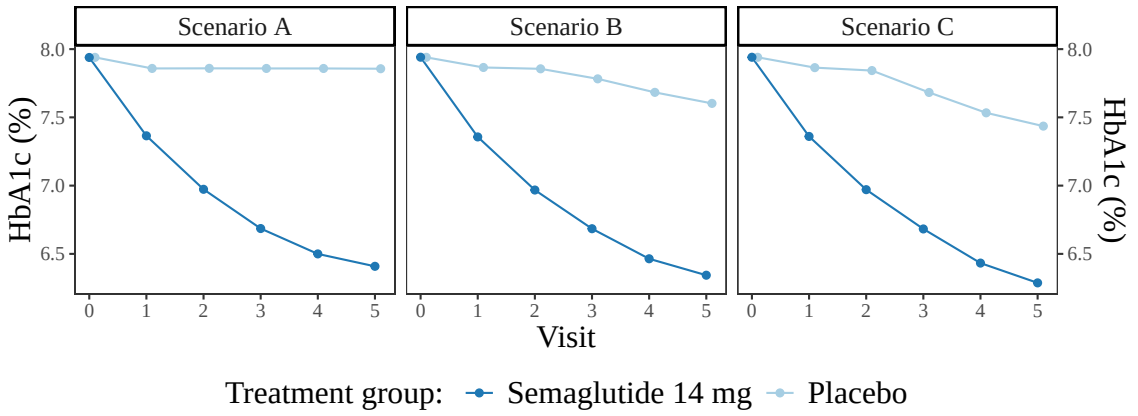


Figure 6.3: *Mean plots for each of the three scenarios.*

6.2 Results

In this section we will compare the performance of the models presented in Chapter 5 on simulated data from the three different scenarios described in the above section. Specifically we will focus on their ability to capture the true treatment effect (6.1.7), when using a hypothetical strategy for ICEs.

To be specific, the different models that we will be comparing in this section are the empirical mean given in (5.3.6), the linear model (5.3.7) and the MMRM given in (5.3.8). In addition to these, we will fit two LTMLE models that differ in their assumptions on the dependencies between the involved variables. The first one will make very few assumption on the variables by taking the dependencies implied by the DAG in Figure 5.1 into account. The second one will make more assumptions on the dependencies as it will only consider the dependencies implied by Figure 5.2. The MMRM is fitted using the R-package `mmrm` [35] and the LTMLE is implemented using the `ltmle` package [36]. For more information on how all these models are implemented, see Section 7.2. In Table 6.3 all the results are summarised.

Table 6.3: *Mean estimate and bias across 1000 simulated datasets, from each of the five different models, in three different rescue scenarios. In addition the RMSE and coverage are reported.*

Scenario	Method	Mean estimate	Mean bias	RMSE	Coverage
A	**Empirical mean**	−1.486	0.01410	0.3467	95.4%
	Linear model	−1.486	0.01385	0.3392	95%
	MMRM	−1.488	0.01151	0.3314	95.2%
	LTMLE (Figure 5.1)	−1.485	0.01509	0.01042	95.1%
	LTMLE (Figure 5.2)	−1.485	0.01477	0.01041	95.1%
B	**Empirical mean**	−0.5204	0.9796	1.038	19.5%
	Linear model	−0.5874	0.9126	0.9728	24.9%
	MMRM	−1.503	−0.002691	0.3545	94%
	LTMLE (Figure 5.1)	−1.479	0.02074	0.01236	94.3%
	LTMLE (Figure 5.2)	−1.481	0.01863	0.01291	94.8%
C	**Empirical mean**	0.1337	1.634	1.676	1.4%
	Linear model	0.03843	1.538	1.585	3.1%
	MMRM	−1.504	−0.004362	0.3949	93%
	LTMLE (Figure 5.1)	−1.480	0.01964	0.01307	90.1%
	LTMLE (Figure 5.2)	−1.496	0.003724	0.01428	93.9%

If we solely consider the data without initiation of rescue medication, scenario A, the estimates across the methods are approximately equal. But both of the proposed LTMLE models have lower RMSE and hence lower variance of the estimate. In scenario B, when there is introduced rescue medication in the data, the empirical mean and the linear model become very biased compared to the remaining models, which is also reflected in their coverage which is far away from 95%. The MMRM slightly underestimates the treatment effect and the LTMLEs slightly overestimate it. However, it is clear that the LTMLE based on more assumptions introduces more bias than the one based on Figure 5.1. Looking towards the situation with even more rescue medication, scenario C, the conclusion is almost the same as for scenario B. Both the empirical mean and the linear model fail to capture the true treatment effect, with horrible confidence intervals. Even when almost 40% of the data is missing, both MMRM and the LTMLEs succeed in determining the treatment effect with low bias and variance. This result really emphasizes the need of models that handle the longitudinal data structure. Across all scenarios it is remarkable that both LTMLEs obtain a much lower RMSE than any of the other methods.

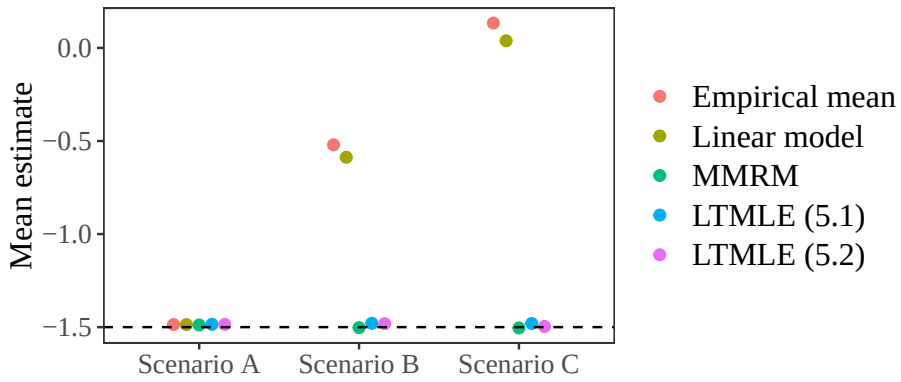


Figure 6.4: *Illustrating the mean estimates across data from different scenarios, corresponding to results from Table 6.3. Dots are slightly jittered in the direction of the x-axis for better visualisation.*

Figure 6.4 displays the mean estimates, which gives a nice overview of the mean bias of the different models across each for the 1000 simulated datasets in each scenario. It is clear that the dots corresponding to MMRM and the LTMLEs remain close to the true value no matter the rescue scenario. However, the empirical mean and linear model are only close to the true estimate in scenario A. For the two other scenarios, the linear model obtains estimates that are slightly closer to the true value than the empirical mean. From this figure it is hard to tell the performance of the MMRM and LTMLEs apart. In Figure 6.5 we have zoomed in and are only considering these methods such that it is easier to compare them. From the first plot it seemed that they were pretty accurate, however all the models show a little bias.The MMRM has the lowest mean bias across all scenarios except scenario C, where the LTMLE based on Figure 5.2 has the lowest mean bias. There is a clear tendency that the LTMLE based on the dependencies in Figure 5.2 has smaller mean bias than the one based on the dependencies in Figure 5.2 across all scenarios. This difference is very small in scenario A and B, but becomes a bit larger in scenario C. We would expect this pattern to be the other way around as more assumptions may introduce more bias if not satisfied.

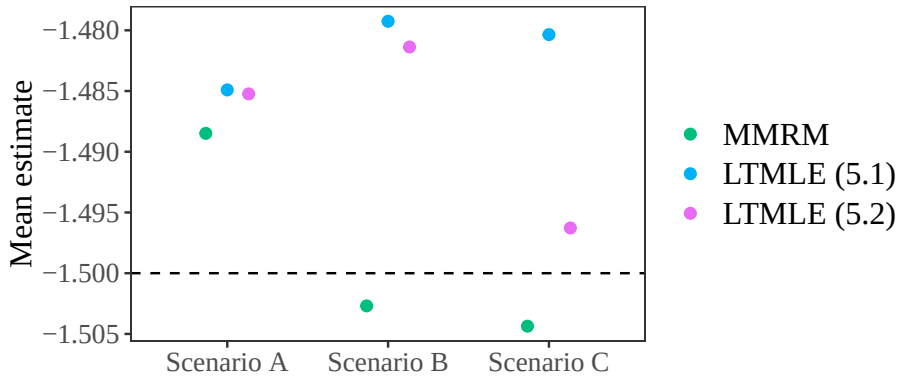


Figure 6.5: *Illustrating the mean estimates across data from different scenarios, corresponding to selected results from Table 6.3. Dots are slightly jittered in the direction of the x-axis for better visualisation.*

6.2.1 Discussion on model assumptions

As we are in complete control of the true DGP for the simulated data in this chapter, we also have complete knowledge on whether or not the assumptions made by each of the methods is actually satisfied. Common for all methods is that they rely on the identifiability assumptions to be able to make a causal interpretation of the results.

Identifiability assumptions

Assumption 5.1 states the assumptions that were needed to identify the causal estimand of interest by the statistical estimand (5.2.3). We will discuss the validity of these assumptions using the DAG presented in Figure 6.1. The sequential randomisation assumption is satisfied as there are no unmeasured confounders in the A_k-Y relationship outside of $W_{0:k}$, $A_{0:(k-1)}$ and $Z_{1:(k-1)}$ for $k = 0, \ldots, 4$. Likewise there are no unmeasured confounders in the Z_k-Y relationship. Moving on to positivity, we know that

$$P(A_0 = 1 \mid W_0) = P(A_0 = 1) = 1/2 \tag{6.2.1}$$

$$P(A_k = 1 \mid W_{0:k}, A_0, A_{1:(k-1)} = \mathbb{1}_{k-1}, Z_{1:(k-1)} = \mathbb{0}_{k-1}) = 0.98 \tag{6.2.2}$$

$$P(Z_k = 0 \mid W_{0:k}, A_0, A_{1:k} = \mathbb{1}_k, Z_{1:(k-1)} = \mathbb{0}_{k-1}) = \text{expit}(\gamma_{0,k} + \gamma_{1,k} A_0 + \gamma_{2,k} W_k) \tag{6.2.3}$$

which is strictly larger than zero for all $k = 1, \ldots, 4$, as expit : $\mathbb{R} \to]0, 1[$. Hence, the simulated data satisfies the positivity assumption. The sequential consistency assumption is also satisfied by construction, see (6.1.6).

Assumptions for the parametric models

The MAR assumptions stated in (5.3.4) requires that there are no confounders in the relationship between the censoring variable R, which is determined by $A_{1:4}$ and $Z_{1:4}$, and the repeated measurements of HbA_{1c} after censoring time.

More intuitively, the MAR assumption is satisfied if the observations after an ICE give no extra information about the probability of censoring, beyond what has happened up until that point in time. That is, the probability of discontinuation at visit k when we know the past and the future, is equal to the probability in the case where we only know the past, just as explained mathematically in Section 5.3. However, as there is a direct arrow from the variables A_k and Z_k to W_{k+1} in the DAG 6.1, we know that having information on the future W_{k+1} will inform about what has happened in the past (A_k and Z_k). Therefore it is hard to imagine that the MAR assumption for censored data after occurrence of ICE should be satisfied in this case.

Both the empirical mean, the linear model and the MMRM assumes MAR. Looking into the results this could be the reason that both empirical mean and the linear model performs bad when we increase the censoring. However, as we see in the results, MMRM is robust against slight deviations in the distribution from normality.

As discussed earlier, common for all parametric models is, that they assume that the relationship between the variables can be described by a finite amount of parameters. This assumption holds true for the simulation study as we have used parametric models to generate data.

In addition, the parametric models assume that the residuals are normally distributed under correct model specification. This assumption could be questioned too, since the error terms are t distributed in the simulation study.

Assumptions for the LTMLE

In this chapter, we have fitted an LTMLE based on the relationships in Figure 5.1 and an LTMLE based on the relationships in Figure 5.2, which implies some conditional independencies. As discussed in Section 5.1 this is a question of bias-variance trade-off. This is also what we saw in the results presented in Table 6.3 as the estimates from the LTMLE based on more assumptions had higher variance than the ones from the LTMLE that made fewer assumptions. Overall, the two fits gave very similar results in every rescue scenario.

7 Case study

In this chapter, we will use the same methods as we considered in Chapter 6, but since this is only one dataset instead of 1000, we are able to concretise it a bit more. First, we will explain how data was prepared in preparation for analysis and then we will go into detail of the specification of each model in R. Lastly, we summarise the results and discuss the differences between the results from each model.

7.1 Data preparation

In this section, we will shortly explain how we have prepared the data prior to analysis. The aim is also to describe how we have combined different datasets to obtain the needed information.

Table 7.1: *Overview of the different important datasets.*

Dataset	Description	Structure
`adsl`	Subject-Level data	One record per subject
`adcm`	Concomitant Medication data	One record per subject per concomitant medication
`adlb`	Laboratory data	One record per subject per parameter per analysis visit

Table 7.1 displays an overview of the different datasets that we have used in the analysis of data and their overall structure in terms of how many records it contains. The `adsl` dataset contains information on the participants treatment allocation and demographic. In addition it contains a variety of flags for analysis eligibility among others and dates of important events like randomisation, first and last exposure to treatment. The `adcm` dataset contains information on what participants have received rescue medication among other concomitant medication and the date of when it was initiated. The `adlb` dataset contains all the biomarker measurements made for each patient at each visit and the corresponding date of the visit.

In a clinical trial, things happen continuously in time, and not exactly the planned number of weeks after initiation of treatment.However, in this project we have focused on methods using a discrete timescale. Hence, to align with the data structure presented in Section 5.1, we need to discretise these variables. How this has been done will be explained in the following.

Initiation of rescue medication, measured in the Z variable, was judged by comparing the dates of the HbA_{1c} measurements for the corresponding visit in the `adlb` data and the date of initiation of rescue medication in the `adcm` data. Of course it is completely possible for participants to take rescue medication multiple times throughout the study, however, we will not be accounting for this and hence the date of initiation of rescue medication that we are using in this comparison will be the date of the first initiation of rescue medication.

Likewise, the indicator of staying on trial product at the k'th visit, A_k, will be 1 if the participant still takes the randomly allocated treatment and 0 otherwise. This will be judged from the dates of the HbA_{1c} measurements for the corresponding visit in the `adlb` data and the date of treatment discontinuation found in the `adsl` data.

Table 7.2: *Number of participants discontinuing treatment product or initiating rescue medication by treatment arm.*

	Semaglutide 14 mg		**Placebo**		**Total**	
Number of participants	175		178		353	
Discontinued treatment at any time	34	(19.4%)	25	(14.0%)	59	(16.7%)
Received rescue medication at any time	< 5	($< 2.9\%$)	27	(15.2%)	< 32	($< 9.1\%$)

Table 7.2 displays the number of participants who end up either discontinuing the trial product or receiving rescue medication at any time. Remark that there may be an overlap between the participants that discontinue treatment and initiate rescue medication as it is completely possible for a participant to experience both these intercurrent events. There are slightly more participants who receive 14 mg of semaglutide and discontinue treatment than participants in the placebo group. In addition, Table 7.2 makes it clear that the number of participants in need of rescue medication is larger in the placebo group than in the treatment group receiving 14 mg of semaglutide. This is of course expected as the placebo group does not receive any anti-diabetic medication.

7.2 Model specifications in R

In this section, we show how to implement the methods described in Section 5.3 and 5.4 in R and obtaining an estimate of the hypothetical estimand (5.2.3). The models handle ICEs in different ways, either consider only Pre-ICE data or consider all collected data, as described in Subsection 5.2.1.

Example 7.1 Empirical mean

To use the empirical mean for estimating the hypothetical estimand, we only consider the final outcome for participants that did not discontinue nor receive rescue medication. There were 129 participants in the placebo group and 140 in the treatment group that followed this regime of interest, so we ended up only taking these 269 participants into account when conducting this analysis.

The empirical mean method is equivalent to fitting a linear model with change in HbA_{1c} from baseline to week 26 as the dependent variable and the only independent variable being the treatment A_0. To calculate the contrast we use the R package `emmeans` [37], where we specify `trt.vs.control` as A_0 because it is the treatment variable.

```r
model <- lm(HBA1CV5 ~ A0, data)
contrast <- emmeans(model, trt.vs.ctrl ~ A0)
confint(contrast)
```

R Output

```
$emmeans
 A0 emmean     SE  df lower.CL upper.CL
  0 -0.347 0.0967 267   -0.538   -0.157
  1 -1.519 0.0928 267   -1.701   -1.336
```

```
Confidence level used: 0.95

$contrasts
 contrast  estimate    SE  df lower.CL upper.CL
 A01 - A00    -1.17 0.134 267    -1.44   -0.907

Confidence level used: 0.95
```

When fitting the linear model in R and calculating the contrast, we get the output displayed above. The first parts are the estimates of the change in HbA_{1c} from baseline to week 26 in the two treatment groups. The second part is the estimate $\bar{Y}_1 - \bar{Y}_0$ of the hypothetical estimand (5.2.3), using the notation introduced in Subsection 5.3.1.

By this model, the expected difference of change from baseline in HbA_{1c} to week 26 between the treatment and placebo group is -1.17 for participants in the target population that do not initiate rescue medication or discontinue randomised treatment. From this result, we see that it is a statistically significant difference, meaning that the empirical mean model concludes that treatment is superior to placebo in treating T2D.

The estimate in Example 7.1, is solely the difference in the mean change from baseline to week 26 in HbA_{1c} between the treatment groups. Based on this example it seems that treatment is better at lowering HbA_{1c} than placebo, but we want to investigate whether or not this is due to a difference in the HbA_{1c} values at baseline. Next step is then to incorporate the baseline value as a covariate in the model.

Example 7.2 Linear model

Just as in Example 7.1, the linear model introduced in Subsection 5.3.1 for participants that did not take rescue medication or discontinued randomised treatment, when modelling change in HbA_{1c} from baseline to week 26. As mentioned, this model also adjusts for selected baseline covariates.

```r
model <- lm(HBA1CV5 ~ HBA1CBL + A0 + REGION, data = data)
contrast <- emmeans(model, trt.vs.ctrl ~ A0, weights = "proportional")
confint(contrast)
```

R Output

```
$emmeans
 A0 emmean     SE  df lower.CL upper.CL
  0 -0.384 0.0890 262   -0.559   -0.209
  1 -1.485 0.0854 262   -1.653   -1.317

Results are averaged over the levels of: REGION
Confidence level used: 0.95

$contrasts
 contrast  estimate    SE  df lower.CL upper.CL
 A01 - A00     -1.1 0.124 262    -1.34   -0.857

Results are averaged over the levels of: REGION
Confidence level used: 0.95
```

As we are now adjusting for the categorical variable `REGION`, we need to specify how averages are taken across the different levels of this variable. We do this by using the weights argument in the `emmeans` call. To get the average treatment effect on a population

level, where the distribution of people from each region is the same as in the observed data, we use proportional weights. This specification influences the estimated marginal means, however the contrast remains the same.

The output when fitting this model in R is very similar to what we saw in Example 7.1. But, as expected, the confidence interval is a bit narrower in this case since some of the variation in the previous example is now explained through the additional covariates. Again we see that the result is significant in favour of the treatment group with the same interpretation as the previous example.

The linear model in Example 7.2 handles missing data in the response variable by just omitting them in the analysis. It implies that no matter how many observations, that are unaffected by ICEs, a participant has, the outcome will be set to missing and will not be included in the analysis if the participant experiences an ICE at any visit. By omitting participants that experience an ICE, we break randomisation and potentially introduce bias as these participants might differ from the ones that do not experience ICEs. Next up is then to make use of all the observations that are unaffected by ICEs. The amount of Pre-ICE data for each visit is summarised in Table 7.3, which exactly is the data that MMRM will consider.

Table 7.3: *Number of participants included in the MMRM analysis.*

	Treatment	**Placebo**	**Total**
Visit 1	178	175	353
Visit 2	163	169	332
Visit 3	155	157	312
Visit 4	144	151	295
Visit 5	140	129	269

Example 7.3 Mixed model for repeated measures

The model that was used in the original analysis for the hypothetical estimand in [32] was the MMRM described in Subsection 5.3.2 and it is implemented in R in the following. This model uses data in long format which is obtained from the wide format data by running the code displayed in R.2. In addition, the HbA_{1c} measurements are set to missing when the participant experiences an ICE.

```
library(mmrm)
model <- mmrm(Upsilon ~ (HBA1CBL + A0 + REGION)*VISIT +
              us(VISIT | USUBJID), datalong)
contrast <- emmeans(model, trt.vs.ctrl ~ A0|VISIT,
                    weights = "proportional")
confint(contrast)
```

R Output

```
$emmeans
VISIT = Visit 5:
 A0  emmean     SE  df lower.CL upper.CL
  0 -0.0647 0.0894 285   -0.241   0.1112
  1 -1.5029 0.0890 275   -1.678  -1.3278

...
```

```
Results are averaged over the levels of: REGION
Confidence level used: 0.95

$contrasts
VISIT = Visit 5:
 contrast  estimate     SE  df lower.CL upper.CL
 A01 - A00   -1.438 0.1265 280   -1.687   -1.189

...

Results are averaged over the levels of: REGION
Confidence level used: 0.95
```

Selected parts of the output is displayed above. The formula used as the first input in `mmrm` corresponds to the model specified in (5.3.8), where `us(VISIT|USUBJID)` specifies the unstructured covariance matrix.

Contrary to the linear model we have now incorporated interactions between variables. When investigating the treatment effect it is important to specify the levels of the categorical variables that we are interested in. This is done by specifying `trt.vs.ctrl ~ A0|VISIT`, since we want the visit specific treatment coefficients. Specifically, the effect at visit 5, which is the change in HbA_{1c} from baseline to week 26. Just as in the linear model, we specify proportional weights for the `emmeans`.

The estimate of the hypothetical estimand (5.2.3), is -1.44, which is larger than what we saw in Example 7.1 and 7.2. The confidence interval is of similar length, however, it also indicates a larger treatment effect. This means that by including the information on the repeated measurements of change from baseline in HbA_{1c} taken throughout the study duration, we gain evidence pointing toward a greater difference in change in HbA_{1c} from baseline to week 26 for the treatment group than the placebo group.

These three examples above concludes the different parametric methods that we will compare. Now we will apply the theory of targeted maximum likelihood estimation on PIONEER 1.

Example 7.4 Longitudinal targeted maximum likelihood estimation

This example will present the implementation of using LTMLE, see Section 5.4, for estimating the hypothetical estimand (5.2.3). This model has been proposed as a competitor for the MMRM model described in Subsection 5.3.2. Just as the MMRM, it also takes the repeated measurements of HbA_{1c} into account, however, it uses a different approach for handling data affected by ICEs, as it utilises all the observed data even after occurrence of an ICE.

```r
library(ltmle)
set.seed(123)
model <- ltmle(data,
               Anodes = c("A0", "A1", "Z1", "A2", "Z2",
                          "A3", "Z3", "A4", "Z4"),
               Lnodes = c("HBA1CV1", "HBA1CV2", "HBA1CV3", "HBA1CV4"),
               Ynodes = "HBA1CV5",
               abar   = list(c(1, 1, 0, 1, 0, 1, 0, 1, 0),
                             c(0, 1, 0, 1, 0, 1, 0, 1, 0)),
               deterministic.g.function = deterministic.g.function,
               SL.library = "default")
summary(model)
```

R Output

```
Treatment Estimate:
   Parameter Estimate:  -1.5156
    Estimated Std Err:  0.073106
              p-value:  <2e-16
    95% Conf Interval: -1.6588, -1.3723

Control Estimate:
   Parameter Estimate:  -0.087003
    Estimated Std Err:  0.10999
              p-value:  <2e-16
    95% Conf Interval: -0.30258, 0.12857

Additive Treatment Effect:
   Parameter Estimate:  -1.4286
    Estimated Std Err:  0.12906
              p-value:  <2e-16
    95% Conf Interval: -1.6815, -1.1756
```

The LTMLE call in R is a bit different than for the models presented previously. We specify time varying treatments as `Anodes`, as we are now accounting for the use of rescue medication and treatment discontinuation throughout the study. The argument `abar` specifies the treatment regimes that we are interested in, which is no rescue medication $Z_{1:4} = \mathbb{0}_4$ and staying on the randomised treatment $A_{1:4} = \mathbb{1}_4$ in either treatment group.

The `Lnodes` argument is the vector of time-dependent covariate nodes, which in this case are the repeated measurements of change from baseline in HbA_{1c} at the visits up to week 26. The `Ynodes` argument specifies the endpoint of interest, which is change from baseline in HbA_{1c} after 26 weeks. The argument `deterministic.g.function` is a function that incorporates the monotone patterns of A_k and Z_k if one discontinues treatment or initiates rescue medication, as this is factual knowledge of the experiment. In addition, we specify in this function that A_0 is randomised with probability $1/2$. The last argument `SL.library` is the method for making the initial estimate needed in Algorithm 4.1 as LTMLE is just repeated use of TMLE. When using "default" as the `SL.library`, the algorithm uses a so-called super learner to make an initial estimate [36]. This is an algorithm that combines different machine learning methods to make an estimate of the quantity of interest. As one of the machine learning algorithms the super learner uses is random forest, it is necessary to specify a seed prior to fitting the model. However, this seed has very little effect on the final estimate. In this project we will not go further into how this works and to learn more about the super learner, we refer to [38].

The output shows estimates of the mean change in HbA1c with corresponding standard error, p-value and confidence interval in the treatment group, the placebo group and the contrast between these two. The contrast is labelled as the Additive Treatment Effect and it corresponds to the estimate of the hypothetical estimand (5.2.3). The model concludes that the estimate of the contrast is significant, just like the previous examples, indicating that the 14 mg of semaglutide is superior to placebo. The output is pretty similar to the one from MMRM, concluding almost the same size of effect along with approximately equal confidence intervals.

As mentioned in Section 5.4, the algorithm needs some information on the dependencies between the variables to guide the fitting process. However, as none has been specified it defaults to assuming that every variable is dependent on all the previously observed variables. Hence the above LTMLE

fit is under the dependencies presented in Figure 5.1. Next up is to consider an LTMLE that is restricted to only account for the dependencies implied by the DAG in Figure 5.2.

Example 7.5
In the following, we present how to implement the LTMLE algorithm to incorporate knowledge on the dependencies between variables. Specifically, we will implement the LTMLE accounting for the dependencies displayed in Figure 5.2. To implement this we will include the arguments `Qform` and `gform` in which we specify the dependencies of variables. We write `Q.kplus1` as the dependent variable in the elements in `Qform` following the syntax of the `ltmle` package, which is in line with the notation in Section 5.4. The syntax for specifying these dependencies is rather similar to specifying the linear models, however, as an important note, this does not imply that LTMLE is parametric. These formulas just specify some dependencies, which is then used to form an initial estimate using a super learner as mentioned above.

```
set.seed(123)
model <- ltmle(data,
               Anodes = c("A0", "A1", "Z1", "A2", "Z2",
                          "A3", "Z3", "A4", "Z4"),
               Lnodes = c("HBA1CV1", "HBA1CV2", "HBA1CV3", "HBA1CV4"),
               Ynodes = "HBA1CV5",
               abar   = list(c(1, 1, 0, 1, 0, 1, 0, 1, 0),
                             c(0, 1, 0, 1, 0, 1, 0, 1, 0)),
               deterministic.g.function = deterministic.g.function,
               SL.library = "default")
               Qform = c(HBA1CV1 = "Q.kplus1 ~ REGION + HBA1CBL + A0",
                         HBA1CV2 = "Q.kplus1 ~ REGION + A0 + HBA1CV1 +
                                    Z1 + A1",
                         HBA1CV3 = "Q.kplus1 ~ REGION + A0 + HBA1CV2 +
                                    Z2 + A2",
                         HBA1CV4 = "Q.kplus1 ~ REGION + A0 + HBA1CV3 +
                                    Z3 + A3",
                         HBA1CV5 = "Q.kplus1 ~ REGION + A0 + HBA1CV4 +
                                    Z4 + A4"),
               gform = c("A0 ~ 1",
                         "A1 ~ HBA1CV1 + A0",
                         "Z1 ~ HBA1CV1 + A0",
                         "A2 ~ HBA1CV2 + Z1 + A1",
                         "Z2 ~ HBA1CV2 + Z1 + A1",
                         "A3 ~ HBA1CV3 + Z2 + A2",
                         "Z3 ~ HBA1CV3 + Z2 + A2",
                         "A4 ~ HBA1CV4 + Z3 + A3",
                         "Z4 ~ HBA1CV4 + Z3 + A3"))
summary(model)
```

R Output

```
Treatment Estimate:
   Parameter Estimate:  -1.5184
    Estimated Std Err:  0.076062
              p-value:  <2e-16
    95% Conf Interval: -1.6675, -1.3693

Control Estimate:
   Parameter Estimate:  -0.025725
    Estimated Std Err:  0.13224
              p-value:  <2e-16
    95% Conf Interval: -0.28491, 0.23346
```

```
Additive Treatment Effect:
   Parameter Estimate:  -1.4926
    Estimated Std Err:  0.15099
              p-value:  <2e-16
    95% Conf Interval: -1.7886, -1.1967
```

As explained above the function will output three different estimates. We are mainly interested in the last estimate, the contrast between the two first estimates, which again is significant, with the same interpretation as the previous example.

7.3 Results

In Table 7.4, the results from each of the models presented in the above examples are summarised. The contrast estimates, presented in the first column are the estimates of the hypothetical estimand (5.2.3). The estimates in the second column, corresponding to the control group, are estimates of the average change from baseline in HbA_{1c} after 26 weeks among the participants taking placebo and following the treatment regime of interest. Likewise, the last column contains estimates of the average change from baseline in HbA_{1c} after 26 weeks among participants taking 14 mg of oral semaglutide.

Table 7.4: *Estimates of the contrast and mean change from baseline to week 26 in HbA_{1c} across the treatment groups and their corresponding 95% confidence intervals according to different models.*

Data used	Method	Semaglutide 14 mg	Placebo	Contrast
Endpoint only	**Empirical mean**	-1.51 $[-1.68, -1.32]$	-0.35 $[-0.55, -0.14]$	-1.17 $[-1.44, -0.91]$
	Linear model	-1.49 $[-1.65, -1.32]$	-0.38 $[-0.56, -0.21]$	-1.1 $[-1.34, -0.86]$
Repeated measures	**MMRM**	-1.50 $[-1.68, -1.33]$	-0.06 $[-0.24, 0.11]$	-1.44 $[-1.69, -1.19]$
	LTMLE (**Figure 5.1**)	-1.52 $[-1.66, -1.37]$	-0.09 $[-0.30, 0.13]$	-1.43 $[-1.68, -1.18]$
	LTMLE (**Figure 5.2**)	-1.52 $[-1.67, -1.37]$	-0.03 $[-0.28, 0.23]$	-1.49 $[-1.79, -1.20]$

Common for all four models is that they agree on the fact that 14 mg of oral semaglutide is significantly better than the placebo group at reducing the HbA_{1c} in T2D patients. MMRM and LTMLE pretty much agree on a larger treatment effect than what the two other methods indicate. Hence all models conclude that 14 mg of semaglutide is a superior treatment to placebo in treating T2D.

To investigate the performance even more, we illustrate the estimates and their corresponding confidence interval in forest plots in Figure 7.1 for both 14 mg semaglutide, placebo and the contrast. This especially highlights that the different models agree on the estimate in the treatment group, and that it is the estimation in the placebo group which make them differ a lot. In the treatment group, we see that the confidence interval becomes notably more narrow from MMRM to both of the LTMLE estimates. This difference is not as remarkable in neither the placebo group not in the contrast estimate.

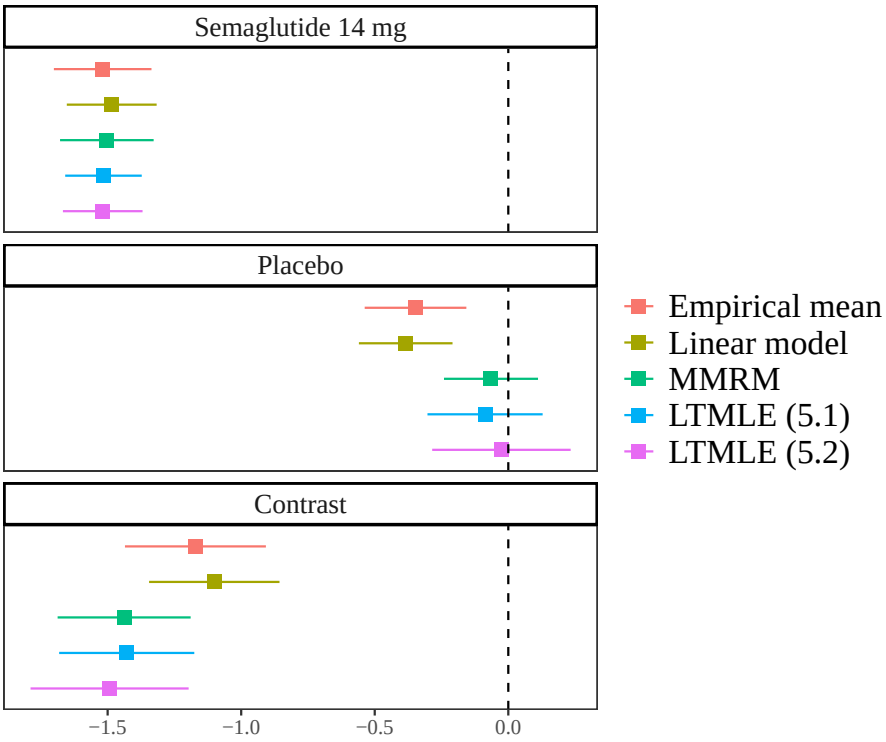


Figure 7.1: *Forest plot of the estimates from each estimation method.*

As mentioned, the methods do pretty much agree on the estimate of change from baseline in the treatment group. However, they do not agree on the estimate of change from baseline in the control group, which is also clear from the forest plot. This discrepancy may be explained by the fact that intake of rescue medication is very unbalanced in the favour of the placebo group, as one can see in Table 7.2. All the participants in the placebo group who experience an ICE are removed from the estimation methods who only take the endpoint into consideration. Hence we are left with all the participants who, contrary to what was expected, lowered their HbA_{1c} on placebo due to diet and exercise. To investigate the difference among the participants that did not take rescue medication and the ones that did, we have displayed the mean HbA_{1c} for these two scenarios across all visits grouped by treatment assignment in Figure 7.2. Here it is clear that the participants that did not end up receiving rescue within either treatment assignment differ from the ones that did by having much lower HbA_{1c} values. It is also clear that there is a tendency for the participants in the placebo arm that do initiate rescue to get higher and higher HbA_{1c} values up until visit 3, where the HbA_{1c} is lowered. This is most likely due to the effect of the rescue medication. We do not see the same tendency for the semaglutide 14 mg arm, however, this is most likely due to the effect of treatment lowering the HbA_{1c} value continuously throughout the study. In contrast, the methods, which accounted for repeated measures of the outcome variable, knew the general trajectory of a participant receiving placebo, at least until the point where they receive rescue medication or discontinued the randomised treatment. Hence it is expected for them to make less biased estimates than the methods that only have information on the final outcome.

As an other important note, there is a large difference in computational burden between the LTMLEs and the parametric models. LTMLE takes approximately 1 minute, which is 60 times longer than the MMRM. This is something that is important to take into consideration when making large simulation studies, as this additional time for each model fit accumulates to a large amount of time.

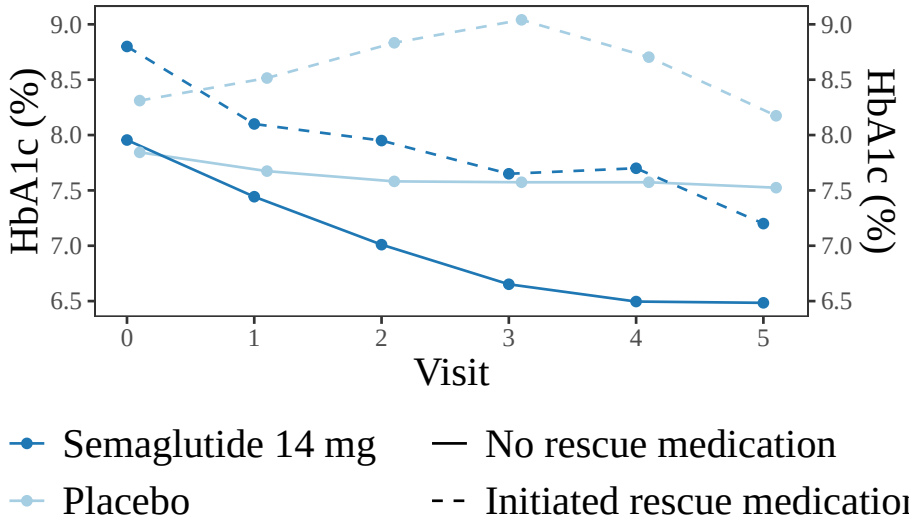


Figure 7.2: *Mean plot for the participants that do and do not initiate rescue medication at some point throughout the study in either treatment arm.*

8 Discussion

When one seeks to do causal inference to investigate effect of treatment, RCTs are the golden standard as some assumptions are automatically fulfilled, causing fewer problems with identification. However, as RCTs often last for a long period of time to prove long-term safety and efficacy, it is common for participants to deviate from the treatment assignment, either in terms of discontinuing randomised treatment or initialising other, possibly disease targeting, medication. ICH introduced the estimand framework, that specifies five attributes clarifying the effect of interest. This is of great importance when working with studies where ICEs occur as ICEs are known to have an effect on the measurements that are used to estimate the treatment effect. This implies that they also affect the interpretation of the estimated effect which highlights the importance of clearly specifying how these are handled in the analyses. We incorporated this estimand framework into a workflow for making causal inference.

The effect of ICEs on the measurements taken in the study is a problem when we are interested in the hypothetical estimand, which is the treatment effect when participants continue their randomised treatment and do not initiate rescue medication. But even though clinical trials aim to be a controlled environment, deviations from the treatment regime of interest are impossible to prevent as they can be a consequence of something external either caused by treatment or something completely unrelated. In this project we discussed different analysis methods that took different amounts of data into account in order to estimate this hypothetical estimand.

In our simulation study we wanted to investigate how the methods handled varying amounts of rescue medication. It showed approximately equal performance across all five proposed methods when participants were not allowed to take rescue medication. However, when introducing rescue medication both the MMRM and the LTMLE methods significantly improves the estimation results in terms of bias and RMSE compared to the two linear regressions, but LTMLE obtained a smaller variance. This could imply differences in performance in smaller samples, which could also be a topic of interest for a future simulation study. Hence, when estimating the hypothetical estimand in cases where data is affected by ICEs, one should always favour models that include repeated measures at least up until the time point where the participant no longer complies with the treatment regime of interest.

The two LTMLEs that were introduced in this project proved to have a lower RMSE than the MMRM. However, some drawbacks of this method are the complex and not very transparent theory behind it and the computational burden. The algorithm takes much longer to arrive at an estimate than the MMRM. This is partly due to the super learner that is used to find an initial estimate. To reduce the time spent finding the initial estimate using the super learner, one could instead use a simpler regression method. However, this might also affect the final estimate, so it is a trade-off. In the end, the LTMLEs and the MMRM were very similar in all other aspects than the RMSE.

In the project, we solely considered static interventions, where all participants are assigned to the

same exact regime of not receiving rescue medication nor discontinuing the randomised treatment. However, completely eliminating the possibility of receiving rescue medication may not be feasible in practice, due to ethical reasons. Hence one could question the need for an estimand that reflects a treatment effect that is not necessarily feasible for all participants in practice. In this case one could look towards "milder" and more realistic interventions allowing some participants to initiate rescue medication, for example only the 5% with worst glycaemic control, implying a stochastic or dynamic intervention. The important thing to remember is the ultimate goal to balance the confounding effect of rescue medication across treatment groups. Dynamic and stochastic interventions open up for this possibility as opposed to static interventions where the only option is to completely eliminate all use of rescue medication.

When fitting the different models on the real data did not highlight other differences than what we saw in the simulation study. We aimed to replicate the original analysis from PIONEER 1, as they also used an MMRM for the hypothetical estimand. However, it should be noted that we used the software R instead of the software SAS for the analyses, which is used in the original analyses from the PIONEER 1 trial. Even though we used different software the results are the same as the ones reported for the hypothetical estimand in the PIONEER 1 trial. In the protocol for PIONEER 1, it is specified that whenever a participant meets the rescue medication threshold, an additional measurement after a few days is required. A consequence of this is measurements at unscheduled visits, which contain more information to whether or not a participant ends up receiving rescue medication. We have focused solely on the scheduled visits in this project, however in [27] they investigated methods which also took the measurements from unscheduled visits into account. They concluded that "*Our simulation study demonstrated that using only all data from regular scheduled visits up to the initiation of rescue medication is inappropriate. This is because the unscheduled confirmatory visit values are ignored so that the MAR assumption is no longer satisfied.*" Hence, one could possibly extend what we did in this project to something that accounts for the additional measurements.

The nonparametric approach using the LTMLE algorithm is limited to pathwise differentiable parameters. However, there are some parameters that might be of interest, which are not pathwise differentiable. For example, the conditional average treatment effect for participants at age 40 is not a pathwise differentiable parameter, since age is a continuous variable. Moreover, we have focused only on one-dimensional parameters in this report, since we are only interested in the treatment effect under some specific intervention. There exist on-going research in extending this algorithm to target infinite dimensional parameters.

9 Conclusion

In this project, we explored the challenges associated with evaluating hypothetical estimands in clinical trials, where ICEs can impact treatment effect assessments. After a presentation of standard methodologies, including linear models of the outcome and the commonly used MMRM, we proposed LTMLE as a more robust alternative for estimating the hypothetical treatment effect. Our simulations and empirical analyses indicate that while MMRM offers an easily interpretable solution, LTMLE provides a more accurate reflection of causal relationships, particularly in scenarios involving ICEs like rescue medication and treatment discontinuation. However, a drawback of using LTMLE is its high computational burden.

In the simulation study, we saw a clear difference in the performance of the different methods. All methods performed rather well when data was affected very little by ICEs. However, when the amount of participants experiencing ICEs increased, the simple methods modelling only the outcome were no longer able to provide reliable estimates. The MMRM and LTMLE methods accounted for varying amounts of repeated measurements of the outcome, still they performed similar in terms of bias and coverage in the first two scenarios. In the last scenario, where there was a large amount of participants that experienced ICEs, the LTMLE making few assumptions on the dependencies had higher bias than MMRM and the other LTMLE that made more assumptions. However, it had the lowest RMSE implying the lowest variance estimate and tightest confidence intervals. However, this method obtained a lower coverage, which is a drawback compared to the other methods when the goal is to determine the efficacy of a drug, where control of the type 1 error is especially important. But it is worth remarking, that even the MMRM, which was used in the submission for PIONEER 1, could not maintain a coverage at 95% when increasing the amount of ICEs. In general, LTMLE significantly outperformed MMRM in terms of variance, which is particularly beneficial in smaller samples. Overall, neither bias nor RMSE from the MMRM and LTMLE are significantly affected even when there are a lot of data being affected by ICEs.

In the case study, approximately 20% of the participants were affected by an ICE and hence we expected LTMLE to perform better in terms of accuracy compared to the MMRM using the knowledge we obtained in the simulation study. However, the LTMLE and MMRM models led to very similar estimates of effects and standard errors. An interesting finding in the case study was that with the clear unbalance in ICEs among treatment groups, all methods agreed on the mean change in the treatment arm. Hence the difference in the contrasts among methods was due to the estimates of the mean in the placebo group. The models that solely modelled the endpoint found that placebo had a significant effect in treating T2D patients, however, this was due to the heavy selection bias implied by only accounting for participants that did not experience ICEs.

Our work demonstrates how the integration of the estimand framework can enhance our understanding of treatment outcomes in clinical trials. These estimands are a key part of the workflow for causal inference which translates clinical questions of interest to statistical estimation problems.

We also emphasise the importance of selecting the appropriate model in the causal and statistical world to obtain valid and reliable estimates of causal quantities.

Overall, this research contributes to a foundation for more informed model selection in clinical trial planning. In addition it highlights the need to continue collecting information on ICEs and developing analytical methods that can address the complexities of ICEs in clinical trials.

Bibliography

[1] Ian R. White, Christina Bamias, Pollyanna Hardy, Stuart Pocock, and Jacquie Warner. Randomized clinical trials with added rescue medication: some approaches to their analysis and interpretation. *Statistics in Medicine*, 2001.

[2] Camila Olarte Parra, Rhian M. Daniel, David Wright, and Jonathan W. Bartlett. Estimating hypothetical estimands with causal inference and missing data estimators in a diabetes trial case study. *Biometrics*, 81(1), 2025. URL https://doi.org/10.1093/biomtc/ujae167.

[3] The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials, 2020. URL https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials-scientific-guideline.

[4] The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH E9 Statistical Principles for Clinical Trials, 1998. URL https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf.

[5] Moritz Pohl, Lukas Baumann, Rouven Behnisch, Marietta Kirchner, Johannes Krisam, and Anja Sander. Estimands — A Basic Element for Clinical Trials: Part 29 of a Series on Evaluation of Scientific Publications. *Deutsches Ärzteblatt International*, 118(51-52):883, 2021.

[6] Novo Nordisk A/S. Redacted CSR: PIONEER 1 – Monotherapy. Efficacy and safety of oral semaglutide versus placebo in subjects with type 2 diabetes mellitus treated with diet and exercise only, 2017. URL https://www.novonordisk-trials.com/study-results.html.

[7] Emily Eyth and Rohan Naik. Hemoglobin A1C. *StatPearls*, 2023. URL https://www.ncbi.nlm.nih.gov/books/NBK549816/.

[8] Zhikai Zheng, Yao Zong, Yiyang Ma, Yucheng Tian, Yidan Pang, Changqing Zhang, and Junjie Gao. Glucagon-like peptide-1 receptor: mechanisms and advances in therapy. *Signal Transduction and Targeted Therapy*, 9(1):234, 2024.

[9] American Diabetes Association. Hyperglycemia (high blood glucose), 2022. URL https://www.diabetes.org/healthy-living/medication-treatments/blood-glucose-testing-and-control/hyperglycemia.

[10] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal inference in statistics*. John Wiley & Sons Ltd, 2016.

[11] Mark J. van der Laan and Sherri Rose. *Targeted Learning - Causal Inference for Observational and Experimental Data*. Springer, 2011.

[12] Alejandro Schuler and Mark J. van der Laan. Introduction to modern causal inference, 2023. URL https://alejandroschuler.github.io/mci/introduction-to-modern-causal-inference.html.

[13] Mohammad Ali Mansournia, Mahyar Etminan, Goodarz Danaei, Jay S. Kaufman, and Gary Collins. Handling time varying confounding in observational research. *BMJ*, 359, 2017. URL https://www.bmj.com/content/359/bmj.j4587.

[14] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning series. The MIT Press, Cambridge, 1st edition, 2017.

[15] Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005. URL https://doi.org/10.1198/016214504000001880.

[16] Anastasios A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer New York, NY, 2006.

[17] Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 76:292–304, February 2022. URL https://arxiv.org/pdf/2107.00681.

[18] Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *CoRR*, v3, 2019. URL http://arxiv.org/abs/1908.08526.

[19] Helene Charlotte Rytgaard. *Targeted causal learning for longitudinal data*. PhD thesis, Section of Biostatistics, University of Copenhagen, January 2020.

[20] Mark J. van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2006.

[21] Thomas P. Hettmansperger, Joseph W. McKean, and Simon J. Sheather. Robust nonparametric methods. *Journal of the American Statistical Association*, 95(452):1308–1312, 2000. URL https://www.jstor.org/stable/2669777.

[22] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[23] Richard D. Gill, Mark J. van der Laan, and Jon A. Wellner. Inefficient estimators of the bivariate survival function for three models. *Annales de l'I.H.P. Probabilités et statistiques*, 1995.

[24] Tom Kennedy. A: Conditional expectations. Course Material, 2007. URL https://math.arizona.edu/~tgk/464_07/cond_exp.pdf.

[25] Mark J. Van der Laan and Sherri Rose. *Targeted learning in data science*. Springer, 2018.

[26] U.S. Department of Health, Food Human Services, and Drug Administration. Guidance for industry: Diabetes mellitus: Developing drugs and therapeutic biologics for treatment and prevention, February 2008. URL https://www.regulations.gov/document/FDA-2008-D-0118-0003. Draft guidance.

[27] Björn Holzhauer, Mouna Akacha, and Georgina Bermann. Choice of estimand and analysis methods in diabetes trials with rescue medication. *Pharmaceutical statistics*, 14(6):433–447, 2015.

[28] Benjamin Cuer et al. Handling informative dropout in longitudinal analysis of health-related quality of life: application of three approaches to data from the esophageal cancer clinical trial. *BMC medical research methodology*, 20:1–13, 2020.

[29] Marie Davidian and Anastasios A. Tsiatis. Statistical Methods for Analysis With Missing Data Lecture Notes, 2015.

[30] Camila Olarte Parra, Rhian M. Daniel, and Jonathan W. Bartlett. Hypothetical estimands in clinical trials: A unification of causal inference and missing data methods. *Statistics in Biopharmaceutical Research*, 15(2):421–432, 2022. URL https://doi.org/10.1080/19466315.2022.2081599.

[31] Melanie L. Bell and Brooke A. Rabe. The mixed model for repeated measures for cluster randomized trials: a simulation study investigating bias and type i error with missing continuous data. *Trials*, 2020. URL https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-020-4114-9.

[32] Vanita R. Aroda, Julio Rosenstock, Yasuo Terauchi, Yuksel Altuntas, Nebojsa M. Lalic, Enrique C. Morales Villegas, Ole K. Jeppesen, Erik Christiansen, Christin L. Hertz, Martin Haluzík, and PIONEER 1 Investigators. Pioneer 1: Randomized clinical trial of the efficacy and safety of oral semaglutide monotherapy in comparison with placebo in patients with type 2 diabetes. *Diabetes Care*, 42(9):1724–1732, June 2019. URL https://doi.org/10.2337/dc19-0749.

[33] The European Medicines Agency (EMA). Guideline on clinical investigation of medicinal products in the treatment or prevention of diabetes mellitus, 2023. URL https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-clinical-investigation-medicinal-products-treatment-or-prevention-diabetes-mellitus-revision-2_en.pdf.

[34] Mark J. van der Laan and Susan Gruber. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics*, 2012. URL https://pubmed.ncbi.nlm.nih.gov/22611591/.

[35] Daniel Sabanes Bove. Package 'mmrm', 2024.

[36] Joshua Schwab et al. Package 'ltmle', 2023.

[37] Russell V. Lenth. Package 'emmeans', 2025.

[38] Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007. URL https://www.degruyter.com/document/doi/10.2202/1544-6115.1309/html.

[39] Douglas A. Wolfe and Grant Schneider. *Primer for data analytics and graduate study in statistics*. Springer, Cham, Switzerland, 1st edition, 2020.

[40] Lokenath Debnath and Piotr Mikusinski. *Introduction to Hilbert Spaces with Applications*. Elsevier Science & Technology, 2005.

Appendices

A Supplementary material

This appendix is a collection of supplementary definitions, results and some proofs that are used in the main part of the project.

A.1 Probability theory

In this section we will list definitions, lemmas and theorems, and append some of them with their respective proof.

Definition A.1 (Convergence in distribution [39])**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of k-dimensional stochastic vectors such that X_n has cumulative distribution function F_n. If X is a continuous stochastic vector with cumulative distribution function F and

$$F_n(x) \to F(x), \text{ as } n \to \infty \text{ for all } x \in \mathbb{R}^k \tag{A.1.1}$$

holds, it is said that $(X_n)_{n\in\mathbb{N}}$ *converges in distribution* to X. We will denote convergence in distribution by $X_n \xrightarrow{\text{d}} X$. In addition, we say that $(X_n)_{n\in\mathbb{N}}$ has an *asymptotic distribution* with cumulative distribution function F.

The notation $\xrightarrow[P]{d}$ is used when it is important to highlight the specific distribution P that it converges with respect to.

Lemma A.2 ([22])**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of stochastic vectors and let X be some stochastic vector. Convergence in distribution $X_n \xrightarrow{d} X$ is equivalent to*

$$E[f(X_n)] \to E[f(X)] \tag{A.1.2}$$

for all bounded, continuous functions f.

Definition A.3 (Convergence in probability [39])**.** Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of stochastic variables. If

$$\lim_{n\to\infty} P(|Y_n - y| > \epsilon) = 0 \tag{A.1.3}$$

for some constant $y \in \mathbb{R}$ and every $\varepsilon > 0$, we say that the sequence $(Y_n)_{n\in\mathbb{N}}$ *converges in probability* to y as $n \to \infty$ and write $Y_n \xrightarrow{P} Y$.

Theorem A.4 (Slutsky's theorem [39]). *Consider two sequences of stochastic variables $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$. Assume that*

$$X_n \xrightarrow{d} X \tag{A.1.4}$$

$$Y_n \xrightarrow{P} y \tag{A.1.5}$$

for some stochastic variable X and constant $y \in \mathbb{R}$. Then

$$X_n + Y_n \xrightarrow{d} X + y. \tag{A.1.6}$$

Theorem A.5 (Central limit theorem [12]). *Let $X_1, X_2 \ldots$ be i.i.d observations of a stochastic variable X that follows a distribution with mean μ and variance σ^2. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then*

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} Z \tag{A.1.7}$$

where $Z \sim \mathcal{N}(0, 1)$.

Sometimes we will abuse notation by writing

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1) \tag{A.1.8}$$

instead of (A.1.7).

Lemma A.6 (Le Cam's third lemma [22]). *Let $(P_n)_{n\in\mathbb{N}}$ and $(Q_n)_{n\in\mathbb{N}}$ be sequences of probability measures on measurable spaces, and let $(X_n)_{n\in\mathbb{N}}$ be a sequence of k-dimensional stochastic vectors. Suppose that Q_n is contiguous with respect to P_n and*

$$\begin{bmatrix} X_n \\ \log \frac{dQ_n}{dP_n} \end{bmatrix} \xrightarrow[P_n]{d} \mathcal{N}_{k+1}\left(\begin{bmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{bmatrix}, \begin{bmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{bmatrix} \right), \tag{A.1.9}$$

then

$$X_n \xrightarrow[Q_n]{d} \mathcal{N}_k(\mu + \tau, \Sigma). \tag{A.1.10}$$

For the definition of contiguity see [22, defn. 6.3].

Theorem A.7 ([24]). *Let X, Y, Z be random variables. Assuming that the expectations exist, the following equality holds*

$$E[E[X \mid Z, Y] \mid Y] = E[X \mid Y]. \tag{A.1.11}$$

Proof. Suppose that the random variables X, Y and Z are discrete. We start by considering the

left hand side of the equation (A.1.11). Given that $Y = y$, the random variable $E[X \mid Z, Y]$ has possible values $E[X \mid Z = z, Y = y]$ with probability $P(Z = z \mid Y = y)$, where z varies across the range of Z. Therefore,

$$\begin{aligned}
E[E[X \mid Z, Y] \mid Y = y] &= \sum_z E[X \mid Z = z, Y = y] P(Z = z \mid Y = y) \\
&= \sum_z \sum_x x P(X = x \mid Z = z, Y = y) P(Z = z \mid Y = y) \\
&= \sum_{z,x} x \frac{P(X = x, Z = z, Y = y)}{P(Z = z, Y = y)} \frac{P(Z = z, Y = y)}{P(Y = y)} \\
&= \sum_{z,x} x \frac{P(X = x, Z = z, Y = y)}{P(Y = y)} \\
&= \sum_x x \frac{P(X = x, Y = y)}{P(Y = y)} \\
&= \sum_x x P(X = x \mid Y = y) \\
&= E[X \mid Y = y],
\end{aligned} \tag{A.1.12}$$

where we have used the definition of the expectation of a discrete variable and the definition of conditional probability multiple times. Moreover, in the fifth equality we used the definition of marginal distributions. The proof is similar for continuous random variables, where the summations are replaced with integrals. □

Theorem A.8. *Assume that X is a continuous random variable, and that A is a binary random variable. Then*

$$E[X \mid A = 1] = \frac{E[I(A = 1)X]}{P(A = 1)}. \tag{A.1.13}$$

Proof. Considering the expectation of X conditional on $A = 1$ we know that

$$\begin{aligned}
E[X \mid A = 1] &= \int x f_{X|A}(x \mid 1) dx \\
&= \int x \frac{f_{X,A}(x, 1)}{P(A = 1)} dx \\
&= \frac{1}{P(A = 1)} \int x f_{X,A}(x, 1) dx \\
&= \frac{1}{P(A = 1)} \sum_{a=0}^{1} \int x I(a = 1) f_{X,A}(x, a) dx \\
&= \frac{E[I(A = 1)X]}{P(A = 1)}.
\end{aligned}$$

□

A.2 The Hilbert space $\mathcal{L}^2$

This section will give a short introduction to the Hilbert space $\mathcal{L}^2$ and an important result in this context. It is based on [40].

Definition A.9. Define $\mathcal{L}^2(P_0)$ as the set of all functions of the random variable $O \sim P_0$ with sample space $\mathcal{O}$, that have finite variance, that is,

$$\mathcal{L}^2(P_0) = \left\{ f(o) : \int_{\mathcal{O}} f(o)^2 dP_0 < \infty \right\}. \tag{A.2.1}$$

Often the probability distribution is left out and $\mathcal{L}^2(P_0)$ will simply be denoted $\mathcal{L}^2$ when the probability distribution is clear from the context.

The space of functions $\mathcal{L}^2$ is an infinite-dimensional Hilbert space, that is, $\mathcal{L}^2$ is

- Linear: $f, g \in \mathcal{L}^2, \alpha \in \mathbb{R} \implies f + \alpha g \in \mathcal{L}^2$.
- An inner product space: $\langle f, g \rangle = E_{P_0}[fg]$, where f and g are said to be *orthogonal* if $E_{P_0}[fg] = 0$.
- A complete space.

Definition A.10. Define $\mathcal{L}_0^2(P_0)$ as the set of all functions of the random variable $O \sim P_0$ with sample space $\mathcal{O}$, that have mean zero and finite variance, that is,

$$\mathcal{L}_0^2(P_0) = \left\{ f(o) : \int_{\mathcal{O}} f(o)^2 dP_0 < \infty \text{ and } E_{P_0}[f(O)] = 0 \right\}. \tag{A.2.2}$$

Theorem A.11 (Projection theorem for Hilbert spaces)**.** *Let $\mathcal{H}$ be a Hilbert space and let $\mathcal{U}$ denote a closed linear subspace. For any $h \in \mathcal{H}$, there exists a unique $u_0 \in \mathcal{U}$ that is closest to h in the following sense:*

$$\|h - u_0\| \leq \|h - u\|, \quad \forall u \in \mathcal{U}. \tag{A.2.3}$$

We refer to u_0 as the projection of h onto $\mathcal{U}$. Furthermore, it holds for all $u \in \mathcal{U}$ that

$$\langle h - u_0, u \rangle = 0. \tag{A.2.4}$$

A.3 Proof of Theorem 3.12

We let $\varepsilon = 1/\sqrt{n}$, hence it is clear that $\varepsilon \in [0, 1]$ for $n \in \mathbb{N}$. Later it will become important that ε depends on n in this way, so keep this in mind throughout the proof. Consider an arbitrary path

$$\left\{ \tilde{P}_\varepsilon : \varepsilon = 1/\sqrt{n}, \tilde{P}_\varepsilon|_{\varepsilon=0} = P \right\} \subset \mathcal{M}, \tag{A.3.1}$$

which is assumed to be differentiable and denote its score by h. The assumption of it being differentiable is a technicality and for a formal definition we refer to [22, p. 362].

As we are considering a differentiable path, we get

$$\log \prod_{i=1}^{n} \frac{d\tilde{P}_\varepsilon}{dP}(O_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(O_i) - \frac{1}{2} E_P\left[h(O)^2\right] + o_P(1) \tag{A.3.2}$$

by [22, Lemma 25.14]. The left hand side is a random variable and at first sight this result might look strange, however, it will serve as an important tool later. We can use CLT to conclude that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(O_i) \xrightarrow[P]{d} \mathcal{N}(0, E_P[h^2]). \tag{A.3.3}$$

Using this together with (A.3.2) and Slutsky's theorem A.4, it now holds that

$$\log \prod_{i=1}^{n} \frac{d\tilde{P}_\varepsilon}{dP}(O_i) \xrightarrow[P]{d} \mathcal{N}\left(-\frac{1}{2} E_P[h^2], E_P[h^2]\right). \tag{A.3.4}$$

Likewise, we use the assumed property of asymptotic linearity to conclude

$$\sqrt{n}\left(\hat{\Psi}_n - \Psi(P)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_P(O_i) + o_P(1) \xrightarrow[P]{d} \mathcal{N}(0, E_P[\phi_P^2]). \tag{A.3.5}$$

Combining these we find that

$$\begin{bmatrix} \sqrt{n}\left(\hat{\Psi}_n - \Psi(P)\right) \\ \log \prod_{i=1}^{n} \frac{d\tilde{P}_\varepsilon^n}{dP^n}(O_i) \end{bmatrix} \xrightarrow[P]{d} \mathcal{N}_2\left(\begin{bmatrix} 0 \\ -\frac{1}{2}E_P[h^2] \end{bmatrix}, \begin{bmatrix} E_P[\phi_P^2] & E_P[h\phi_P] \\ E_P[h\phi_P] & E_P[h^2] \end{bmatrix}\right), \tag{A.3.6}$$

as $n \to \infty$. Then we can utilize Lemma A.6, which tells us that

$$\sqrt{n}\left(\hat{\tilde{\Psi}}_n - \Psi(P)\right) \xrightarrow[\tilde{P}_\varepsilon]{d} \mathcal{N}\left(E_P[h\phi_P], E_P[\phi_P^2]\right). \tag{A.3.7}$$

Note that the distribution changes and so does the estimator sequence, which is now a collection of estimators based on n draws from $\tilde{P}_\varepsilon$.

With the help of these technical results, we have been able to use asymptotic linearity to say something about the difference between the estimate, as we change the underlying distribution from which samples are taken, and the truth at P. Note that in (A.3.7) we actually see the same variance as in (A.3.5), where we stay in the true distribution P and sample from this. When adding zero, subtracting the mean and moving some terms around in (A.3.7) we find that as $n \to \infty$,

$$\begin{aligned} &\sqrt{n}\left(\hat{\tilde{\Psi}}_n - \Psi(\tilde{P}_\varepsilon) + \Psi(\tilde{P}_\varepsilon) - \Psi(P)\right) - E_P[h\phi_P] \\ &= \sqrt{n}\left(\hat{\tilde{\Psi}}_n - \Psi(\tilde{P}_\varepsilon)\right) + \left(\sqrt{n}\left(\Psi(\tilde{P}_\varepsilon) - \Psi(P)\right) - E_P[h\phi_P]\right) \xrightarrow[\tilde{P}_\varepsilon]{d} \mathcal{N}\left(0, E_P[\phi_P^2]\right). \end{aligned} \tag{A.3.8}$$

Now, recall the regularity property (3.1.28) for an asymptotically normal estimator

$$\sqrt{n}\left(\hat{\tilde{\Psi}}_n - \Psi(\tilde{P}_\varepsilon)\right) \xrightarrow[\tilde{P}_\varepsilon]{d} \mathcal{N}\left(0, E_P[\phi_P^2]\right). \tag{A.3.9}$$

Note that this looks similar to (A.3.8). Since we have assumed both asymptotic linearity and regularity, (A.3.8) and (A.3.9) must hold at the same time, implying that

$$\lim_{n\to\infty} \sqrt{n}\left(\Psi(\tilde{P}_\varepsilon) - \Psi(P)\right) = E_P[h\phi_P], \tag{A.3.10}$$

must hold for the excess in (A.3.8) to be zero. Now, recall that $\varepsilon = 1/\sqrt{n}$, then

$$\lim_{n\to\infty} \sqrt{n}\left(\Psi(\tilde{P}_\varepsilon) - \Psi(P)\right) = \lim_{\varepsilon\to 0} \frac{\Psi(\tilde{P}_\varepsilon) - \Psi(P)}{\varepsilon} \tag{A.3.11}$$

which then implies exactly what we wanted to show. As the h is simply the score of an arbitrary path starting in P, (3.2.2) holds for any path starting in P, when considering a RAL estimator. □

A.4 Identification

In this section we will prove identification of the causal estimand when we have two visits after baseline, that is, we have the data structure $O = (W_0, A_0, Z_0, W_1, A_1, Z_1, Y)$ with intervention variables A_0, A_1, Z_0 and Z_1. Let $Y(a_0, a_1, z_0, z_1)$ be the notation for the potential outcome at the end of the trial when $A_0 = a_0$, $A_1 = a_1$, $Z_0 = z_0$ and $Z_1 = z_1$. Consider the causal estimand of interest

$$\Psi_1^*(P^*) = E_{P^*}[Y(1,1,0,0)], \tag{A.4.1}$$

which reflects the situation of adherence and no rescue medication while receiving active treatment. That this causal estimand can by identified by a statistical estimand is proven in the following, where the equalities rely on Theorem A.7 and Assumption 5.1.

$$\begin{aligned}
&\Psi_1^*(P^*) \\
&= E[Y(1,1,0,0)] \\
&= E\Big[E[Y(1,1,0,0) \mid W_0]\Big] \\
&= E\Big[E[Y(1,1,0,0) \mid W_0, A_0 = 1]\Big] \\
&= E\Big[E[Y(1,1,0,0) \mid W_0, A_0 = 1, Z_0 = 0]\Big] \\
&= E\Big[E\Big[E[Y(1,1,0,0) \mid W_0, A_0 = 1, Z_0 = 0, W_1] \;\Big|\; W_0, A_0 = 1, Z_0 = 0\Big]\Big] \\
&= E\Big[E\Big[E[Y(1,1,0,0) \mid W_0, A_0 = 1, Z_0 = 0, W_1, A_1 = 1] \;\Big|\; W_0, A_0 = 1, Z_0 = 0\Big]\Big] \\
&= E\Big[E\Big[E[Y(1,1,0,0) \mid W_0, A_0 = 1, Z_0 = 0, W_1, A_1 = 1, Z_1 = 0] \;\Big|\; W_0, A_0 = 1, Z_0 = 0\Big]\Big] \\
&= E\Big[E\Big[E[Y \mid W_0, A_0 = 1, Z_0 = 0, W_1, A_1 = 1, Z_1 = 0] \;\Big|\; W_0, A_0 = 1, Z_0 = 0\Big]\Big] \\
&= \Psi_1(P).
\end{aligned}$$

The sequential randomisation assumption ensures the third, fourth, sixth and seventh equality. The second and fifth equality are consequences of Theorem A.7, and lastly the eighth equality is valid due to the sequential consistency assumption. The sequential positivity assumption guarantees that the conditional expectations are well-defined.

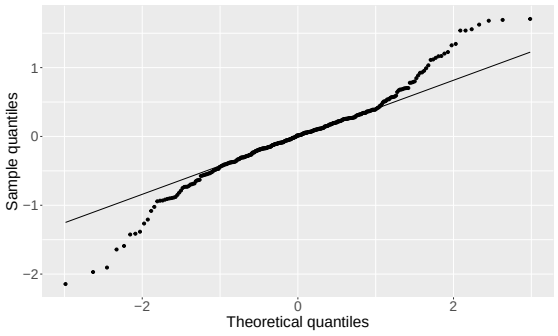
B Overview of clinical trial data

B.1 Inclusion and exclusion criteria for PIONEER 1

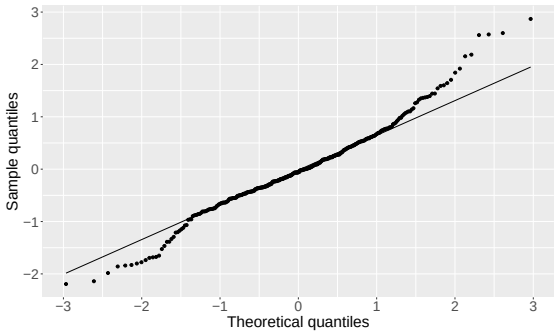
Table B.1: *Inclusion and exclusion criteria for PIONEER 1 [6, sec. 9.3].*

Inclusion criteria	**Exclusion criteria**
Informed consent obtained before any trial-related activities. Trial-related activities are any procedures that are carried out as part of the trial, including activities to determine suitability for the trial.	Known or suspected hypersensitivity to trial product(s) or related products.
Age $\geq$ 18 years. Age $\geq$ 19 years for Algeria only. Age $\geq$ 20 years for Japan only.	Previous participation in this trial. Participation is defined as signed informed consent.
Type 2 diabetes diagnosed clinically $\geq$ 30 days at time of screening.	Female who is pregnant, breast-feeding or intends to become pregnant or is of child-bearing potential and not using an adequate contraceptive method.
Treatment with diet and exercise for at least 30 days prior to day of screening.	Receipt of any investigational medicinal product within 90 days before screening.
HbA_{1c} is between $7.0 - 9.5\%$ at the screening visit.	Any disorder, which in the investigator's opinion might jeopardise subject's safety or compliance with the protocol.
	Family or personal history of multiple endocrine neoplasia type 2 (MEN 2) or medullary thyroid carcinomas (MTC).
	History of pancreatitis (acute or chronic).
	History of major surgical procedures involving the stomach potentially affecting absorption of trial product (e.g. subtotal and total gastrectomy, sleeve gastrectomy, gastric bypass surgery).
	Any of the following: myocardial infarction, stroke or hospitalisation for unstable angina or transient ischaemic attack within the past 180 days prior to the day of screening and randomisation.
	Subjects presently classified as being in New York Heart Association (NYHA) Class IV.
	Planned coronary, carotid or peripheral artery revascularisation known on the day of screening.
	Subjects with alanine aminotransferase (ALT) > 2.5 x upper normal limit (UNL).
	Renal impairment defined as estimated glomerular filtration rate (eGFR) <60 mL/min/1.73 m2 as per Chronic Kidney Disease Epidemiology Collaboration formula (CKD-EPI).
	Treatment with any medication for the indication of diabetes or obesity in a period of 90 days before the day of screening. An exception is short-term insulin treatment for acute illness for a total of $\leq$ 14 days.
	Proliferative retinopathy or maculopathy requiring acute treatment. Verified by fundus photography or dilated fundoscopy performed within 90 days prior to randomisation.
	History or presence of malignant neoplasms within the last 5 years (except basal and squamous cell skin cancer and in-situ carcinomas).

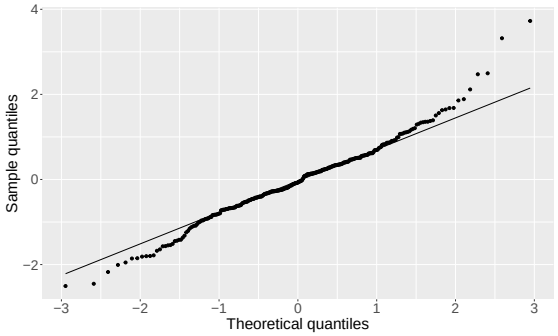
B.2 Normal Quantile-Quantile plots



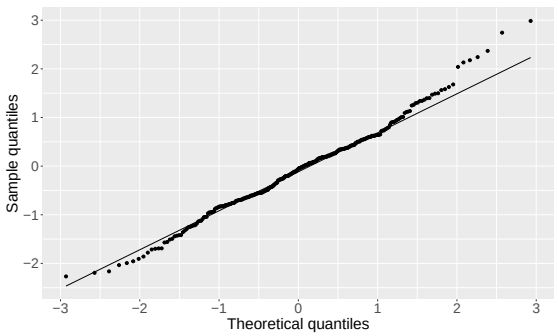
(a) *Visit 1 (week 4)*



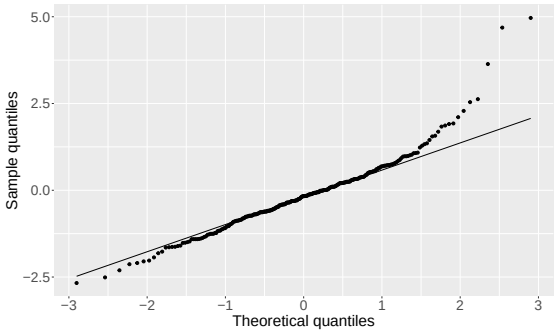
(b) *Visit 2 (week 8)*



(c) *Visit 3 (week 14)*



(d) *Visit 4 (week 20)*



(e) *Visit 5 (week 26)*

Figure B.1: *Normal Quantile-Quantile plot for the residuals from the MMRM presented in Subsection 5.3.2.*

B.3 Distribution of data from PIONEER 1

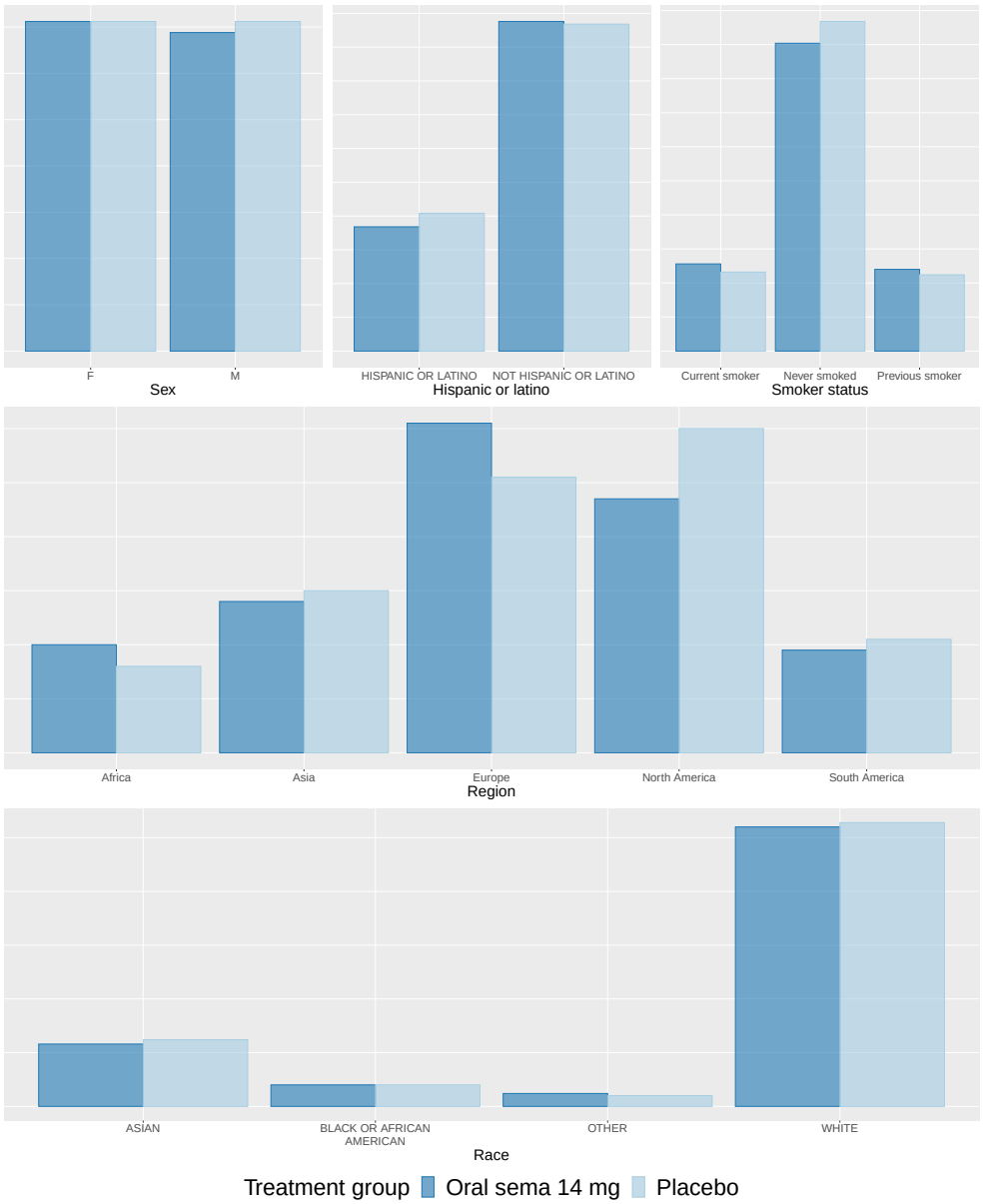


Figure B.2: *Empirical distribution of selected categorical covariates. To protect personal information, the race category labelled OTHER includes not only the category itself but also categories with a few number of participants.*

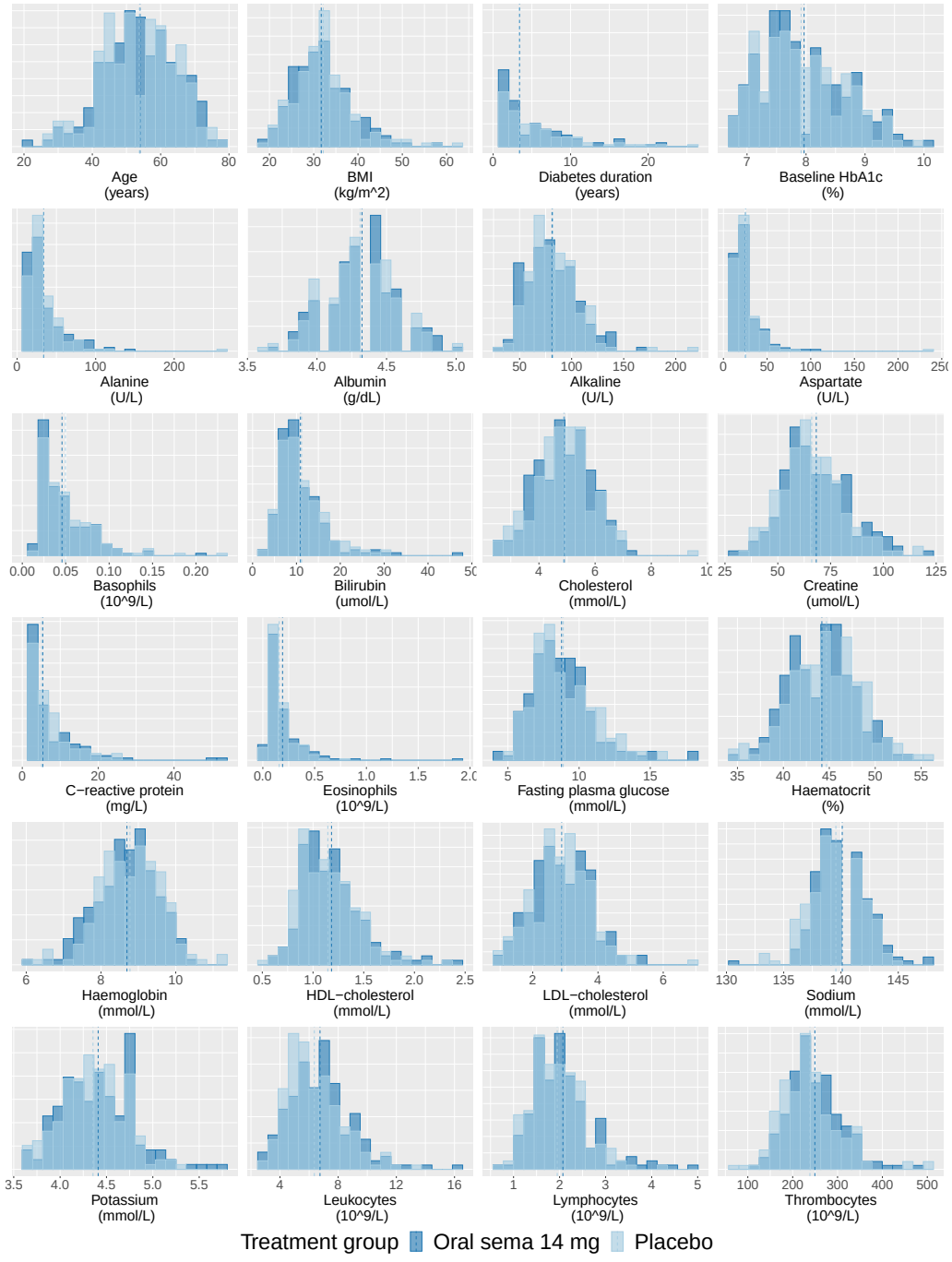


Figure B.3: *Empirical distribution of selected continuous covariates. The means are illustrated by the dashed lines.*

R Code

R.1 Generation of simulated data

```r
generate.data <- function(theta1 = -0.5,
                          theta2 = -0.7,
                          N.covs = 3,
                          coefs = c(-0.05, 0.1, 0.05),
                          cov = diag(1, N.covs),
                          df = 4,
                          N.control = 200,
                          N.treatment = N.control,
                          N.sim = 1000,
                          N.cores = 20,
                          N.visits = 5, # max 5
                          disc.prob = rep(0.02, N.visits),
                          resc = 1){
  if (N.covs < 1) {stop("N.covs must be at least 1")}
  if (length(disc.prob) != N.visits){
    stop("Vector containing probabilities of discontinuation must have length
         equal the number of visits")}
  out <- list()

  # Create variable names for the baseline covariates
  varnames <- paste0("x", 1:(N.covs))

  if (!is.null(N.control) & !is.null(N.treatment)) {
    # For every simulation we create one dataset
    out <- mclapply(1:N.sim, function(k){
      # Create a data frame with the id variable
      id <- data.frame("USUBJID" = 1:(N.control + N.treatment) %>% as.factor())
      reg <- data.frame("REGION" =  sample(c(0:5,4,5), N.control + N.treatment,
                                           replace = TRUE) %>% as.factor())

      # Generate the baseline covariates
      cov[1,1] <- sd(real_data$HBA1CBL)**2
      data <- mvrnorm(n = N.control + N.treatment,
                      mu = c(mean(real_data$HBA1CBL), rep(0, N.covs - 1)),
                      Sigma = cov) %>% as.data.frame()
      colnames(data) <- varnames

      # Combine the id variable with the covariates
      data <- cbind(id, reg, data)

      # Create the treatment variable
      data$A0 <- c(rep(0, N.control), rep(1, N.treatment)) %>% factor()

      # Creating variable such that we don't add multiple rescue effects
```

```
      data["sumZ"] = rep(0, N.control + N.treatment)
      # Creating variable that makes sure we don't stop treatment twice
      data["sumA"] = rep(1, N.control + N.treatment)

      # Create time varying variables for each visit
      for (i in 1:N.visits) {
        var <- paste0("HBA1CV", i)
        if (i == 1) {
          data[[var]] <- theta1*(N.visits - i + 1)/N.visits*(data$A0 == "1") +
            (data$REGION == "3")*0.5 + rt(N.control + N.treatment, df)

          X <- model.matrix(formula(
            paste0("HBA1CV", i, " ~ ",
                   paste0("(",paste0("x", 1:N.covs, collapse = "+"),")^2"),
                   "+",paste0("I(x", 1:N.covs, "^2)", collapse = "+"))),
            data = data %>% mutate_at("x1",~(scale(.))%>%as.vector))[,-c(1,2)]

          data[[var]] <- data[[var]] + data$x1 +
            X %*% c(rep(coefs[2], N.covs-1),
                    rep(coefs[1], N.covs),
                    rep(coefs[3], max(0, N.covs)),
                    rep(coefs[1], max(0, ncol(X) - 3*N.covs)))
          data[[var]] <- data[[var]][,1]

          # Creating the adherence variable
          avar <- paste0("A", i)
          data[[avar]] <-
            case_when(data[["sumA"]] == 1 ~
                        as.integer(rbinom(n = N.control + N.treatment,
                                          size = 1,
                                          p = 1 - disc.prob[i])),
                      TRUE ~ 0)

          data[["sumA"]] <- data[["sumA"]] + 1 - data[[avar]]
          data[[avar]] <- factor(data[[avar]], levels = c("1","0"))

          # Z1 variable
          model1 <- glm(Z1 ~ HBA1CV1 + A0,
                        family = binomial(link = 'logit'),
                        data = real_data %>%
                          mutate(A0 = factor(A0, levels = c("1","0"))))

          data$Z1 <-
            as.integer(rbinom(n = N.control + N.treatment,
                              size = 1,
                              p = pmin(resc*predict(model1,
                                                    newdata = data,
                                                    type = "response"),1)))

          data$sumZ <- data$sumZ + data$Z1

        }else{data[[var]] = data[[paste0("HBA1CV", i - 1)]] +
          theta1*(N.visits - i + 1)/
          N.visits*(data$A0 == "1")*(data[[paste0("A", i - 1)]] == "1") +
          theta2*data[[paste0("Z", i - 1)]] + rt(N.control + N.treatment, df)
        }
        if (i != N.visits & i != 1){
          # Creating the adherence variable
          avar <- paste0("A", i)
          data[[avar]] <-
            case_when(data[["sumA"]] == 1 ~
```

```
                        as.integer(rbinom(n = N.control + N.treatment,
                                          size = 1,
                                          p = 1 - disc.prob[i])),
                                   TRUE ~ 0)

          data[["sumA"]] <- data[["sumA"]] + 1 - data[[avar]]
          data[[avar]] <- factor(data[[avar]], levels = c("1","0"))

          # Creating the rescue medication variable
          model <- glm(paste0("Z",i, " ~ ", " A0 + HBA1CV", i),
                       family = binomial(link = 'logit'),
                       data = real_data %>%
                         mutate(A0 = factor(A0, levels = c("1","0"))))

          zvar <- paste0("Z", i)
          data[[zvar]] <-
            case_when(data[["sumZ"]] == 0 ~
                        as.integer(
                          rbinom(n = N.control + N.treatment,
                                 size = 1,
                                 p = pmin(resc*predict(model,
                                                       newdata = data,
                                                       type = "response"),1))),
                      TRUE ~ 0)

          data[["sumZ"]] <- data[["sumZ"]] + data[[zvar]]
        }
      }
      data %>% rename(HBA1CBL = x1) %>% select(-sumZ, -sumA)
    }, mc.cores = N.cores)

  }
  attr(out, "ATE") <- calc_ATE(N.visits, theta1)
  out
}
```

R.2 From wide to long format

```
datalong <- data %>%
    mutate(HBA1CV2 = case_when((A1 == 0 | Z1 == 1) ~ NA, TRUE ~ HBA1CV2),
           HBA1CV3 = case_when((is.na(HBA1CV2) | A2 == 0 | Z2 == 1) ~ NA,
                               TRUE ~ HBA1CV3),
           HBA1CV4 = case_when((is.na(HBA1CV3) | A3 == 0 | Z3 == 1) ~ NA,
                               TRUE ~ HBA1CV4),
           HBA1CV5 = case_when((is.na(HBA1CV4) | A4 == 0 | Z4 == 1) ~ NA,
                               TRUE ~ HBA1CV5)) %>%
    pivot_longer(cols = c(HBA1CV1, HBA1CV2, HBA1CV3, HBA1CV4, HBA1CV5),
                 names_to = "VISIT", values_to = c("Upsilon")) %>%
    mutate(VISIT = factor(case_when(VISIT == "HBA1CV1" ~ "Visit 1",
                                    VISIT == "HBA1CV2" ~ "Visit 2",
                                    VISIT == "HBA1CV3" ~ "Visit 3",
                                    VISIT == "HBA1CV4" ~ "Visit 4",
                                    VISIT == "HBA1CV5" ~ "Visit 5"))) %>%
    select(USUBJID, HBA1CBL, REGION, A0, VISIT, Upsilon)
```