Comparison of causal inference methods for observational data with a hierarchical structure

ANDRIANA KOSTOURAKI

Department of Non-Communicable Disease Epidemiology
Faculty of Epidemiology and Population Health
LONDON SCHOOL OF HYGIENE AND TROPICAL MEDICINE

Thesis submitted in accordance with the requirements for the degree of
Doctor of Philosophy of the University of London
Funded by the Cancer Research UK
Affiliation with the Inequalities in Cancer Outcomes research group

University of London **December 2024**

To living

Declaration

I, Andriana Kostouraki, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. Chapter 2 was heavily based on the content of the report submitted in partial fulfilment of the requirement for upgrading from MPhil to PhD at the School, which is stored in the LSHTM database.

ANDRIANA KOSTOURAKI
December 2024

Supervisors

Aurélien Belot, Associate Professor in Medical Statistics, London School of Hygiene and Tropical Medicine
Clémence Leyrat, Associate Professor in Statistics, London School of Hygiene and Tropical Medicine
Bernard Rachet, Professor of Cancer Epidemiology, London School of Hygiene and Tropical Medicine

Advisory Committee Member

Miguel Angel Luque-Fernandez, Honorary Associate Professor, London School of Hygiene and Tropical Medicine

Funding

Andriana Kostouraki was supported by the Cancer Research UK Doctoral Studentship in Inequalities in Cancer Outcomes (CR UK Programme C7923/A30945). This work was conducted as part of the corresponding CR UK Programme of the Inequalities in Cancer Outcomes (ICON) group; more details on the related research of the group can be found at https://icon.lshtm.ac.uk/.

Acknowledgements

A huge thank you should go to my three supervisors: Aurélien Belot, Clémence Leyrat and Bernard Rachet, for their continuous patience, guidance and tolerance. This narrative would have never been told without their immense amount of support. Next, I would like to thank David Hajage, Elizabeth (Fizz) Williamson and Guillaume Chauvet, for their collaboration that produced my very first publication. David put an extreme amount of work, as he performed all the simulations that provided key findings for the paper introduced in Chapter 5 of this thesis. Guillaume kindly helped with the mathematical aspects of the paper and, Fizz very kindly contributed her propensity score expertise, valuable advice and (crucial for the shape of the paper) ideas. I would also like to thank my fellow PhD students, and especially Suzanne Keddie, for having our digital 'Newbie Coffee' sessions, which helped me significantly, particularly at the very early stages of my PhD that started remotely. I am also very grateful to Jenny Fleming and Lauren Dalton, for their timely answers to any of my many emails throughout my studies, as well as to Penny Bloore, who has helped with important administrative aspects, especially closer to the ending of my studentship. Additionally, every ICON member (Inequalities in Cancer Outcomes Network) deserves a big thanks for the ideas, kind attitude and subtle way of helping throughout my learning journey. However, a special thanks should go to: Yuki Alencar, for organizing all of these very fun ICON meet-ups, Matthew James Smith, for the valuable 'Causal Inference Journal Club' sessions and Quentin Rollet, for very helpful informal discussions about clustering and various other topics! Lastly and more importantly, I want to thank my parents, for bearing with me, over the course of my studies.

Abstract

Randomised controlled trials (RCTs) are the gold standard for estimating causal treatment effects. When RCTs are not feasible or ethical, causal questions can be answered through observational data. However, analyses of observational data are prone to confounding bias and adequate statistical methods (under explicit assumptions) are needed for addressing it.

Causal inference methods have been proposed to estimate treatment effects when covariates, treatment and outcome are defined at the same unit level (single-level setting). These can be broadly classified into those focusing on the outcome mechanism (e.g., g-computation), those modelling the treatment assignment (e.g., propensity scores) and those combining both (doubly robust methods). In practice, research questions could be about a system-level characteristic (e.g., general versus specialized hospital) or a patient-level characteristic (e.g., type of surgery), while patients are nested within hospital, geographic region, etc. Only few studies have extended causal techniques to hierarchical data, but they were not empirically evaluated for binary outcomes.

In this thesis, we began with a review of three different causal methods, namely, g-computation, inverse probability-of-treatment weighting (IPTW) and augmented inverse probability-of-treatment weighting. We next extended them to hierarchical settings and evaluated their performance under different scenarios of unmeasured cluster-level confounding via Monte Carlo simulations, for the estimation of the population averaged treatment effect with treatment assigned at the individual level. We then derived unified formulas for IPTW variance estimators when targeting different populations like the treated or the overlap population, first starting within the single-level setting. These approximate closed formulas will form the basis of analytical extensions to the hierarchical setting, as an alternative to the bootstrap. To conclude, although common practice, ignoring clustering in causal analysis of observational studies can lead to incorrect inferences, so we extended tools to this setting and provided practical recommendations to applied researchers.

Table of contents

List of figures

List of tables

ACRONYMS & ABBREVIATIONS

AIPTW Augmented Inverse Probability-of-Treatment Weighting *or* Augmented Inverse Probability-of-Treatment Weights *or* Augmented Inverse Probability-of-Treatment Weighted

ATE Average Treatment Effect

ATO Average Treatment effect in the Overlap

ATT Average Treatment effect in the Treated

CI Confidence Interval

CRT Cluster Randomised Trial

DR Doubly Robust

EIC Efficient Influence Curve

HPC High Performance Computer

HR Hazard Ratio

ICU Intensive Care Unit

IPTW-RA Inverse Probability-of-Treatment Weighting - Regression Adjustment

IPTW Inverse Probability-of-Treatment Weighting *or* Inverse Probability-of-Treatment Weights *or* Inverse Probability-of-Treatment Weighted

MCMC Markov Chain Monte Carlo

MSM Marginal Structural Model

PS Propensity Score

RCT Randomised Controlled Trial

SE Standard Error

SMD Standardised Mean Difference

TMLE Targeted Maximum Likelihood Estimation or Targeted Maximum Likelihood Estimator or Targeted Maximum Likelihood Estimate

Chapter 1

Introduction

1.1 Background

An often overarching aim in epidemiological studies is to calculate the causal effect of an exposure or treatment on an outcome. Since this is normally not possible in practice, one may approximate - i.e. estimate - that effect from a sample, which is assumed representative of the target population. The gold standard for performing such procedures is Randomised Controlled Trials (RCTs) in which the treatment is assigned at random. In this thesis, any exposure of interest (e.g., pharmacological interventions, environmental exposures, etc.) is broadly characterized as *the treatment*. Observational studies are an invaluable tool to draw inferences - unfortunately, the usual challenge of which is to capture confounding; that is, ensuring that we have a sufficient set of variables to remove spurious (non-causal) effects between the treatment and the outcome. Confounders can be defined as variables that affect both treatment and outcome and, if unobserved and not accounted for in the analysis, may bias (sometimes, greatly) the effect estimates (see Figure 1.1) [33, 63].

In practice, an additional challenge is the existence of clustering - i.e. individuals may be nested within clusters (e.g., patients nested within geographic region or hospital), which often creates correlation between individual outcomes of the same cluster. Within cluster, treatment levels could be either the same [38] or varying. We narrow our focus to the latter, which is very common in cancer epidemiology. In this context, confounding may be present at the cluster level (i.e., confounders measured at the level of the cluster - e.g., hospital characteristics), on top of the individual level (i.e., confounders measured at the level of the individual - e.g., patient characteristics). Herein, *we focus on settings where some cluster-level confounders may be unobserved* [5]. In observational settings, no residual confounding is a strong assumption to make at either level. However, having unmeasured variables at the cluster level is an additional concern in settings with more than one level [66].

Nevertheless, investigating impact of unmeasured individual-level confounding, although not our focus herein, is of great importance and has been studied in the literature (e.g., [43]). Electronic Health Records (EHR) provide patient characteristics in great abundance, while characteristics at the cluster, e.g. hospital or NHS Trust level may not always be provided by the data at hand. For example, if the cluster is the hospital information may be limited to the identification code of the hospital where the patient was hospitalized, but no further information on specific hospital characteristics may be provided or maybe access to specific hospital characteristics is limited. For instance if the question is whether the receipt of a PET-CT scan affects the receipt of surgery [14], availability of PET-CT scans within the diagnostic hospital may be a variable that affects both the treatment and the outcome, but may not be available in the data. In a Canadian case study to estimate the net population reduction in the number of adverse events if the performance of hospitals was improved to specific standards, information on specific hospital characteristics was not available in the data [6]. In a study examining if the assessment by a Lung Cancer Nurse Specialist is associated with receipt of anticancer therapy, there was a need to account for clustering by English Regional Cancer Network [82]. However, clustering was treated as a nuisance in that particular context.

In the field of cancer epidemiology, clustering is a feature that appears very often [67]. The clinical relevance of accounting for clustering is very strong, especially in studies that aim to understand and reduce observed inequalities in cancer outcomes, as patients are treated in different hospitals nested in different socioeconomic areas [14] - which is the applied field that motivated the objectives of this thesis. Firstly, accounting for clustering may add precision to the effect estimates, as not accounting for it may result in erroneous standard errors and confidence intervals [66]. Secondly, it allows quantification of between cluster heterogeneity, which could partly explain observed inequalities in cancer responses.

Causal inference methods have been applied to estimate causal effects of treatments on outcomes in observational studies [30] while taking into account the presence of confounding to obtain unbiased effect estimates. Herein, we focus on analysis methods that require any potential confounder is measured and included in the analysis to provide valid inferences - which excludes classes of methods such as instrumental variable methods [4]. These can be grouped into three categories: (i) those modeling the outcome mechanism, namely *G-computation* or regression adjustment, (ii) those modeling the treatment assignment mechanism, namely, *Propensity Score (PS)* methods and, (iii) those modeling both. The last are called *Doubly Robust (DR)* as long as they provide consistent estimates when at least one of the two models (treatment or outcome) is correctly specified [12]. These approaches

are extensively applied to evaluate causal effects in observational studies in the single-level setting, i.e., when all variables are measured at the same unit (e.g., patient) level.

Regarding targets of inference, the one of primary interest is usually the average treatment effect in *the overall population*, e.g., cancer patients in hospitals across the UK. Nonetheless, as we will present hereafter, sometimes other populations may be of interest too: *the population of the treated* and *those who have equal chance of receiving any of the defined treatment levels*. When clustering is involved, the overall population may be defined either within or across clusters, which in turn determines a different target (or estimand) of inference. Definition of estimands and their interpretation within the hierarchical setting is also a topic of interest within this work, as within the literature it is not always quite explicit what the target of inference was [34].

To account for clustering, a class of models called *mixed or random*[1] *effects models* that consider the heterogeneity between clusters - originating both from unmeasured cluster-level variables and inherent clustering - is widely applied [66, 67]. Nonetheless, when combined with causal inference methods, guidelines for appropriate practice in the epidemiological literature are relatively scarce [16] compared to other disciplines (e.g., econometrics [66]). Another method which accounts for clustering to obtain correct point estimates and corresponding 95% confidence intervals is marginal models and generalized estimating equations (GEE) [56]. However, these are less suitable for studies investigating inequalities, as instead of modeling clustering, they incorporate it as a nuisance parameter to obtain valid statistical inference.

To reiterate, in many observational studies, there is clustering. In cancer research that aims to reduce inequalities it is very important to account for clustering, both to explain heterogeneity between clusters and understand different aspects of observed inequalities in cancer outcomes. In this context, causal inference methods are not that well developed or used in practice, so we aim to fill that gap by investigating related assumptions, estimands and methods. Our focus lies in how different causal inference methods perform in settings of unobserved cluster-level confounding. Since the broad way of grouping these is whether they model the outcome, treatment, or both, we focused on *G-computation (and, in particular, Standardization [43]), Inverse Probability-of-Treatment Weighting (IPTW) and Augmented Inverse Probability-of-Treatment Weighting (AIPTW)* with the use of *mixed effects models* to account for clustering. Next, we present our overarching aim and specific objectives.

[1] we use the two terms interchangeably hereafter.

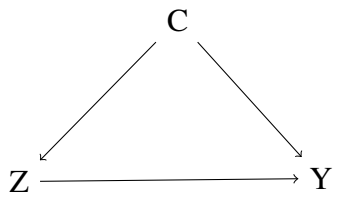


Fig. 1.1 DAG (Directed Acyclic Graph): Z assumed treatment, Y outcome and C confounder.

1.2 Aim and objectives

Our overarching aim is:

To provide recommendations to applied researchers for the implementation of statistical methods for causal inference in the presence of clustering. In particular, to draw conclusions on the performance of G-computation, IPTW and AIPTW estimators, when applied to two-level observational data and treatment is assigned at the individual level. Additionally, to present approaches of estimating the variance of these estimators, to obtain valid 95% confidence intervals and correct statistical inference. The latter, being performed for the single-level setting, will naturally consist the milestone to extend it to the two-level setting in future research. We focus on binary outcomes, as these are very common in practice. We illustrate the behaviour of the estimators of focus: (i) *under different simulation settings and* (ii) *via an application to a dataset of non-small cell lung cancer (NSCLC) patients within the UK - specifically, the example data set represents data from National Cancer Registry at the Office for National Statistics. We present mathematical demonstrations for the derived variance formulas in the single-level setting.*

Therefore, the particular objectives of this thesis are to:

1. Review causal inference methods in the single- and two-level (hierarchical) setting;
2. Clarify assumptions and estimands in these settings;
3. Evaluate and compare the different methods under different scenarios of simulation studies and draw conclusions on their behaviour - especially in presence of unobserved cluster-level confounding - for G-computation combined with mixed effects modeling;
4. Evaluate and compare the different methods under different scenarios of simulation studies and draw conclusions on their behaviour - especially in presence of unobserved cluster-level confounding - for IPTW and AIPTW estimators combined with mixed effects modeling;

5. Investigate the variance estimation methods for G-computation, IPTW and AIPTW estimators in the single-level setting and provide unifying closed-form approximate marginal variance formulas for IPTW estimators, for different target populations; this should be the milestone for future research in the hierarchical setting.

1.3 Outline of the thesis

We now present a short outline for the following chapters of the dissertation:

- Chapter 2 briefly presents the potential outcomes framework and the statistical methods involved to estimate the causal effect of a treatment on an outcome. We start by introducing all the required notation and assumptions that will be used throughout the thesis. We continue by reviewing briefly the literature of causal inference methods, first in the single-level setting, with the objective to transition to the hierarchical (two-level) setting in Chapters 3 and 4.

- Chapter 3 introduces G-computation applied within the two-level framework. In a separate section, we present a related *short communication draft*, with intent to be submitted for consideration for publication.

- Chapter 4 introduces IPTW and AIPTW applied within the two-level framework. In a separate section, we present a simulation study to evaluate and compare the performance of the different estimators presented in Chapters 3 and 4, under various data generating mechanisms.

- Chapter 5 presents three different approaches to estimate the variance when applying IPTW in single-level data, while targeting various different populations, and modeling the treatment assignment with several different models. A *paper published in Statistics in Medicine*, is also presented in a separate section of the chapter.

- Chapter 6 ends with a discussion of findings, conclusions and general implications to statistical analyses when targeting causal effects in hierarchical (two-level) settings. We also discuss potential avenues of future research that may include extensions of estimators - such as the variance estimator, presented in Chapter 5.

- Finally, in the supplementary appendices the reader may find:

  – Theoretical notes and additional simulation results for Research Paper I, in Appendix A;

- Theoretical notes and additional simulation results for Research Paper II, in Appendix B;
- Additional figures for bias results of simulation study of section 4.4, in Appendix C;
- Supplementary notes on the causal assumptions of positivity and consistency, as well as the estimators discussed in the single-level setting, in Appendices D to G.

Chapter 2

Potential outcomes framework and statistical methods for the estimation of causal effects

2.1 Notation, estimands and assumptions

2.1.1 Notation

Although there are different approaches introduced in the literature [23], throughout this dissertation we follow the potential outcomes framework, often attributed both to Neyman [80] and Rubin [74], having been formalized by the latter, for the non-randomised setting. A historical overview of the framework can be found in [31]. We now present some notation. Imagine an observational cohort study consisting of n individuals. Assume an outcome of interest Y, a binary treatment Z, and $\mathbf{X}$ a set of pretreatment characteristics, postulated as confounders. We denote $Y_i(0)$ the outcome of the i^{th} patient were they not to receive the treatment (i.e. Z=0), and $Y_i(1)$ were they to receive the treatment (i.e., Z=1). Under Rubin's potential outcome framework, it is not feasible to observe both of these two hypothetical outcomes at the same time for the same individual (this is why some refer to the term counterfactuals [41] - although we retain the term *potential outcomes* herein). We denote the expected value of $Y(0)$ as $\mathsf{E}[Y(0)]$ and of $Y(1)$, as $\mathsf{E}[Y(1)]$.

2.1.2 Estimands

We broadly call an *estimand*, the quantity we target to estimate and make inferences on - this corresponds to the population of focus.

Throughout this thesis, we primarily focus on the *overall population of patients* and the corresponding effect on that population, namely, the *average treatment effect (ATE)*. Other populations of interest may be: *the population of the treated*, which corresponds to the *average treatment effect in the treated (ATT)*, and, *the population of those with equal chances of receiving any of the defined treatment levels (either treatment or no treatment in our settings)*, which corresponds to the *average treatment effect in the overlap population (ATO)*. Some of the characteristics and desired properties of these populations, as well as the application settings where these are usually encountered, are briefly discussed in Section 2.3 of Paper II, which can be found in section 5.2 of the thesis.

We may go a step further and define the mean of the target population under treatment as: $m_1 = \frac{\mathsf{E}[Y(1)h(x)]}{\mathsf{E}[h(x)]}$, where $h(:)$ is a prespecified function of $\mathbf{x}$, specific to the population of interest [26]. Let us now provide some intuition. Firstly, let us define the probability of receiving treatment conditional on individual characteristics x_i, as $e_i = e(\mathbf{x}_i) = Pr(Z_i = 1 j \mathbf{X}_i = \mathbf{x}_i)$ (more on that in 2.14). When we target the overall population, we aim to estimate the mean outcome value as if everyone were treated (and respectively, untreated), i.e.: $\mathsf{E}[Y(1)]$ (and respectively, $\mathsf{E}[Y(0)]$). In this case, $h(x)$ is equal to 1. To target the population of the treated, we need to multiply the distribution of the original population by the individual probability to be assigned the treatment, given their characteristics, i.e., e_i. For the overlap population, we target those who are equally likely to be assigned either treatment level, hence, $h(x)$ may be equal to the product of both the probabilities to be assigned and not be assigned the treatment - i.e., $e_i(1 \quad e_i)$. An equivalent population, called matching population, may be targeted when multiplying by $min(e_i; 1 \quad e_i)$ [55] (although, generally, $min(e_i; 1 \quad e_i) \leq e_i(1 \quad e_i)$). Both of the last two estimands assign greater weight to individual causal effects when the probabilities of being treated/untreated are equal and equal to 0:5. Mathematically, we may write:

$$h(x) = \begin{cases} 1 & \text{if target is the overall population,} \\ e_i & \text{if target is the population of the treated,} \\ e_i(1 \quad e_i) & \text{if target is the overlap population,} \\ min(e_i; 1 \quad e_i) & \text{if target is the matching population} \end{cases} \tag{2.1}$$

The mean of the target population under no treatment may be defined as well: $m_0 = \frac{\mathsf{E}[Y(0)h(x)]}{\mathsf{E}[h(x)]}$.

Once having obtained the mean outcome values for the target population under treatment and respectively under no treatment, the effect can be expressed by contrasts, depending on the nature of the outcome (e.g., for a binary outcome: as risk difference, logarithm of the

marginal risk ratio or logarithm of the marginal odds ratio). The ATE is a contrast often expressed by the difference in the two means, i.e., $\mathsf{E}[Y(1)] \quad \mathsf{E}[Y(0)]$. Likewise, the ATT:

$$\mathsf{E}[Y(1)jZ = 1] \quad \mathsf{E}[Y(0)jZ = 1] = \frac{\mathsf{E}[Y(1)e_i]}{\mathsf{E}[e_i]} \quad \frac{\mathsf{E}[Y(0)e_i]}{\mathsf{E}[e_i]} \tag{2.2}$$

The ATO and the overlap population:

$$\frac{\mathsf{E}[Y(1)e_i(1 \quad e_i)]}{\mathsf{E}[e_i(1 \quad e_i)]} \quad \frac{\mathsf{E}[Y(0)e_i(1 \quad e_i)]}{\mathsf{E}[e_i(1 \quad e_i)]} \tag{2.3}$$

And, the ATO and the matching population:

$$\frac{\mathsf{E}[Y(1)min(e_i; 1 \quad e_i)]}{\mathsf{E}[min(e_i; 1 \quad e_i)]} \quad \frac{\mathsf{E}[Y(0)min(e_i; 1 \quad e_i)]}{\mathsf{E}[min(e_i; 1 \quad e_i)]} \tag{2.4}$$

It can be proven that sample estimators that target the above estimands are consistent [26, 45]. However, this is beyond the scope of this thesis.

After introducing the required assumptions for the ATE to be identifiable in the single level setting, referred to as identifiability assumptions, we describe the statistical methods of causal inference when targeting the ATE. We note that depending on the target estimand, some of the identifiability assumptions may be relaxed, as explained in 5.2.

2.1.3 Identifiability assumptions

Prior to estimation, to ensure the *causal* ATE of the target population can be identified from observational data, the *identifiability assumptions* must be met. These assumptions cannot be tested and are the following:

1. Consistency, i.e., for the patients that actually received treatment Z of level z, their observed outcome is the same as what it would have been had they received treatment level z via the hypothetical intervention we have in mind [20], [90], [64]:

$$Z_i = z \;\;) \;\; Y_i = Y_i(z) \tag{2.5}$$

   In other words, the consistency assumption is equivalent to simultaneously requiring that: (i) there are no multiple versions of treatment, and, (ii) observing is the same as imposing.

2. No interference, i.e., the potential outcome $Y_i(z)$ for individual i, is independent of the treatment level assigned to any other individual j, with $i \neq j$, for $i = 1; 2; :::; n$ [21].

3. Positivity, i.e., every patient has non zero probability of receiving treatment of level z, for every different level z of treatment Z or, in other words, the observed treatment levels must vary within potential confounders strata [92]:

$$Pr(X = x) > 0 \;\;) \;\; 0 < Pr(Z = z|X = x) < 1; \forall z; x \tag{2.6}$$

   We note that positivity violations may occur in the specific sample due to random sampling or could be existent by design (structural or theoretical positivity). Details on checks for plausibility of positivity violations can be found in the corresponding appendix E.

4. Unconfoundedness or conditional exchangeability or ignorability, or no unmeasured confounding, i.e., the potential outcomes $Y(z)$ are conditionally independent of the treatment Z given the pre-treatment measured covariates X for every treatment level z [32]. Herein, we adopt the term unconfoundedness, to contrast it to *latent unconfoundedness*, a term mentioned by some authors [18] to describe an aspect of the unconfoundedness assumption within the hierarchical setting; this is introduced in 2.3.1:

$$Y(z) \perp\!\!\!\perp Z|X; \forall z \tag{2.7}$$

   Where, the symbol $\perp\!\!\!\perp$ indicates statistical independence. Assuming no unmeasured confounding is equivalent to saying that treatment is randomly allocated within each strata of the observed confounders X. In a RCT, due to randomisation of treatment, treatment groups are on average exchangeable by design (i.e. if in theory we exchanged the labels of the treatment assigned to each individual, we would still obtain the same treatment effect estimate).

We note that often the assumptions of consistency and no interference are presented in combination, under the headline of *SUTVA (Stable Unit Treatment Value Assumption)*. Along with the aforementioned conditions, no measurement error is assumed. Additionally when parametric modeling is involved, correct model specification is required. We note that in RCTs positivity and exchangeability hold by design, whereas in observational studies the validity of these assumptions cannot be reassured. Finally, when targeting populations other than the overall population, some of these assumptions are fairly relaxed. For more on this, the reader may be directed to Section 2.3 of Paper II in Section 5.2 of the thesis.

2.2 Statistical methods for the estimation of causal effects

In observational studies, it often happens to have a systematic difference in the distribution of certain patient characteristics (e.g. age) between the different treatment levels. If these characteristics affect the individual outcome values too, it leads to confounding.

Statistical methods in causal inference are offered to eliminate confounding either by modeling the outcome, the treatment or both, when targeting marginal effects. The first class of methods is called G-computation estimators and they adjust for confounding by eliminating any effect of patient characteristics on the outcome. The second class that model the treatment (or, more specifically, the probability to receive treatment given covariates), eliminates any chance or systematic imbalances of patient characteristics between treatment groups. These are broadly labeled as propensity score (PS) methods. The last category that models both, are called doubly robust (DR), in that they provide consistent estimates if either the outcome or treatment model is correctly specified. By consistent, we mean estimators for which the estimates come closer to the true mean as the sample size increases and at the same time they achieve the smallest variance.

One may be interested in estimating either a marginal effect, a conditional effect (on covariates and/or clusters, e.g., see 3.1), or both. Marginal effects are of primary interest when evaluating effectiveness of an intervention in public health research, as for public health policy making, the interest lies within the effect of moving from control to treatment in the overall population. Nevertheless, one may be additionally interested in the effect within specific individual characteristics or within cluster (e.g., when it is known by subject-matter knowledge that the individual is less likely to move to a different cluster, e.g., remain within the same hospital, school or family, etc.).

Before continuing with a brief description of the aforementioned methods, we provide some intuition for each. Recall that the objective is to estimate $\mathsf{E}[Y(1)] \quad \mathsf{E}[Y(0)]$. We may express this quantity via three equivalent routes:

$$E[E(YjX;Z = 1)] \quad E[E(YjX;Z = 0)] \tag{2.8}$$

or,

$$E[\frac{ZY}{e(X)}] \quad E[\frac{(1 \quad Z)Y}{1 \quad e(X)}] \tag{2.9}$$

or,

$$E[\frac{ZY}{e(X)} \quad \frac{Z \quad e(x)}{e(x)} E_x[E(YjX;Z = 1)]] \quad E[\frac{(1 \quad Z)Y}{1 \quad e(X)} \quad \frac{(1 \quad Z) \quad (1 \quad e(x))}{1 \quad e(x)} E_x[E(YjX;Z = 0)]] \tag{2.10}$$

Which motivates standardization (2.2.1), IPTW (2.2.2) - a particular PS method and, AIPTW (2.2.3), respectively.

2.2.1 Standardization and G-computation

As already mentioned, one way to account for (or, *adjust for*) confounding, is to model the outcome given a set of measured covariates, which are assumed to be confounders [43]. For example, if the outcome is surgery (yes or no), one could model the probability to receive surgery conditional on covariates (e.g., patient characteristics). However, that would target the treatment effect conditional on covariates. To target a marginal treatment effect instead (i.e., not conditional on given patient population strata, but rather marginal across the population of patients), we need an extra step, called *standardization*, which is a specific case of a more general method, called G-computation[1] - this, includes the following:

i. Model the outcome as a function of the treatment and the confounders.

ii. Predict the outcome in the whole population as if everyone were treated (respectively, untreated).

iii. Average these predictions for the treated (respectively for the untreated).

iv. Contrast these averages according to the effect measure of interest (e.g. risk difference).

In this way, we obtain the so-called standardized mean under treatment:

$$\mathsf{E}[Y(1)] = \sum_{x} \mathsf{E}[Y|Z = 1; X = x] \quad \Pr[X = x] \tag{2.11}$$

and, analogously, under no treatment:

$$\mathsf{E}[Y(0)] = \sum_{x} \mathsf{E}[Y|Z = 0; X = x] \quad \Pr[X = x] \tag{2.12}$$

and normally, the ATE:

$$ATE = \mathsf{E}[Y(1)] \quad \mathsf{E}[Y(0)] \tag{2.13}$$

We note that if the set of confounders in the vector of potential confounders, X, in (2.11) and (2.12) are continuous, then one may replace $Pr[X = x]$ with the probability density function $f_X(x)$, and the sum with an integral. To target the *ATT*, one could replace the

[1] Standardization and IPTW are estimators of the g-formula, which is a general method for causal inference introduced by Robins in 1986 [70, 43].

multiplier of the marginal prevalence of the observed characteristics, $Pr[X = x]$, with the conditional prevalence in the treated, $Pr[X = xjZ = 1]$. One way to obtain the empirical variance estimates for the ATE, is the bootstrap. The basic idea of this technique is briefly described in 5.1.3.

2.2.2 Propensity Score methods: Inverse Probability-of-Treatment Weighting

IPTW belongs to a class of estimators called propensity score (PS) estimators (see [72, 94] and Chapter 12 of [43]). The PS was introduced by Rosenbaum and Rubin in 1983 [73]. Supposing a binary treatment, the PS for an individual i, e_i, is the probability of receiving treatment conditional on measured pretreatment characteristics X_i :

$$e_i = e(\mathbf{x}_i) = \Pr(Z_i = 1 | X_i = x_i) \tag{2.14}$$

The aim of the PS is to balance patients' characteristics between treatment groups on their measured individual characteristics X and therefore remove any confounding bias stemming from the measured confounders X - e.g., see Figure 1.1 (DAG), where the aim is to remove the arrow from C to Z, such that characteristic C no longer confounds the observed relationship between Z and Y. This is feasible due to the key balancing property of the PS: the potential outcomes are independent of the treatment assignment conditional on the PS - which, makes it a balancing score [2]. It has been proven that a sufficient vector of measured confounders $\mathbf{x}$ is the 'finest' balancing score; moreover, as long as the PS adjusts for all of $\mathbf{x}$, it is the crudest form of a balancing score, and any score finer than the PS is a balancing score also [73]. Since the true value of the PS in (2.14) is not known in observational studies, we need to estimate it. As long as the sample size is large enough, the balancing property of the true PS is transferable to the estimated PS from the sample, and balance is obtained within the specific sample. A usual practice is to apply a regression model, which is the practice we as well follow throughout this thesis. Other practices of prediction for the PS growing in popularity, include machine learning techniques [49, 76, 59]. Hereafter, we denote as $\hat{e}_i := \hat{e}_i(\mathbf{x}_i; \hat{a})$ the estimated PS for individual i, with covariate values, $\mathbf{x}_i$, and some parameter estimates $\hat{a}$, obtained from an assumed regression model. After the estimation of the PS (see Chapter 4 for modeling choices in the hierarchical setting), one may balance the distributions of the potential confounders between the treated and the untreated by weighting on functions of the PS, matching on the PS, stratifying or adjusting on the PS. For a concise review of

[2] a balancing score is any function of the covariates that provides conditional independence between the potential outcomes and the treatment values if it is conditioned on.

stratification, regression adjustment and matching on the PS, the reader may be directed to [93]. Herein we focus on IPTW as it has desired properties in comparison to the rest PS methods. PS matching does not use the full sample and is harder to perform when targeting the ATE, while stratification is reported to sometimes lead to more biased estimates compared to weighted estimators [57]. Finally, covariate adjustment using the PS does not separate the design from the analysis stage [7].

IPTW is a PS estimator which entails two steps:

i. Estimation of the PS values.

ii. Derivation of the estimates for the estimand of interest (e.g., ATE).

The IPTW estimators for the marginal mean under treatment, $m_1 = \mathsf{E}[Y(1)]$ and under no treatment, $m_0 = \mathsf{E}[Y(0)]$ can be derived:

$$\hat{m}_1 = \left(\sum_{i=1}^{n} \frac{Y_i Z_i}{\hat{e}_i} \right) \Bigg/ \left(\sum_{i=1}^{n} \frac{Z_i}{\hat{e}_i} \right) \tag{2.15}$$

and,

$$\hat{m}_0 = \left(\sum_{i=1}^{n} \frac{Y_i(1 - Z_i)}{(1 - \hat{e}_i)} \right) \Bigg/ \left(\sum_{i=1}^{n} \frac{(1 - Z_i)}{(1 - \hat{e}_i)} \right) \tag{2.16}$$

Estimators (2.15) and (2.16) can be motivated by the corresponding expressions in (4.1) for m_1 and m_0, respectively.

Where the weight for individual i is the inverse probability-of-treatment that was actually received by individual i. Correspondingly, the ATE estimate will be:

$$\hat{ATE} = \hat{m}_1 - \hat{m}_0 \tag{2.17}$$

Statistical inference and 95% confidence intervals (CIs) can be calculated via *i:* a robust (sandwich) variance estimator that *does not* take into account the uncertainty in the PS estimates, *ii:* a robust (sandwich) variance estimator that *does* take into account the uncertainty in the PS estimates, or, *iii:* the bootstrap. The last two options provide a correct nominal coverage rate and correct CIs regardless of the target of inference (i.e., ATE, ATT, ATO), as they *do* account for the fact that the true individual PS values were not known, but rather estimated from the PS model. This is explained in detail in Chapter 5 and the related Research Paper II. More details for the two-step procedure of the IPTW estimation can be found in Appendix E.

We note that when no models are involved, the standardized mean equals the IPTW mean. However, when models are involved, in presence of positivity violations, standardization may provide smaller standard errors, compared to IPTW; nonetheless, this may be over the expense of bias [43]. When targeting populations not strictly defined, such as the overlap population, there is no equivalent to standardization - as there is, for IPTW - see overlap or matching weights, that are used to target that population [52, 55]. Therefore, guidelines are to apply both, as each method may highlight different aspects of the analysis - e.g., possible limitations, due to the sample we have at hand, and so forth. Whenever possible, it is advised to use doubly robust methods that combine models for the treatment and the outcome in the same estimator. This is the focus of the following subsection. The performance of augmented inverse probability-of-treatment weighting - a particular doubly robust method, widely applied within the single-level setting - compared to standardization and IPTW is explored within simulation scenarios for the two-level setting in Chapter 4.

2.2.3 Doubly Robust methods

Doubly robust (DR) estimation merges a model that describes the outcome distribution with one that describes the treatment allocation mechanism (i.e., PS). Either of these models can produce unbiased causal effect estimates individually under correct model specification - in addition to the *identifiability assumptions* of section 2.1.3. The asset of DR estimation is that it provides a consistent estimator (i.e., asymptotically unbiased, and, with variance approximating zero as sample size increases to infinity), if either the outcome or the PS model is correctly specified. This makes it more *robust* to model misspecification in comparison to the IPTW or standardization when applied separately. In the following paragraphs, we present three different DR approaches: IPTW combined with regression adjustment (IPTW-RA), the augmented inverse probability-of-treatment weighting (AIPTW), which essentially includes the estimators of sections 2.2.1 and 2.2.2, and, targeted maximum likelihood estimation (TMLE). Nonetheless, our focus will be on AIPTW when extending to two levels in Chapter 4.

Inverse probability-of-treatment weighting and regression adjustment (IPTW-RA)

IPTW-RA combines outcome regression with treatment modeling. It consists of the following steps: Firstly, one models the treatment allocation mechanism against the possible confounders (i.e., PS model) and derives the weights (IPTW). Secondly, one models the outcome against the postulated confounders and treatment, using the IPTWs derived from the first step as weights. Thirdly, one predicts the expected response if the whole population

were to be treated and untreated, respectively. Lastly, one follows the standardization step, such as in regular standardization. In this way, one obtains the estimates for m_1 and m_0, and correspondingly for the difference between them. SEs and CIs are generated via the bootstrap [84].

Augmented inverse probability-of-treatment weighting (AIPTW)

AIPTW is a DR estimator originally introduced by Robins and colleagues [71]. In general, IPTW may be variable in the case of extreme values of the weights. AIPTW is more robust to model mis-specification and less variable than the IPTW. The AIPTW estimator is constructed by including the IPTW and standardization model into one. To motivate its construction, recall the expression of the ATE in (2.10). The intuition of the AIPTW estimator for the estimation of $m_1 = \mathsf{E}[Y(1)]$ is briefly presented below. The estimation of $m_0 = \mathsf{E}[Y(0)]$ would be of the same logic, and, afterwards the difference in means due to treatment would provide the ATE.

Let us suppose an observational study of n individuals, an outcome Y, a binary treatment Z and some measured predictors X. Based on the idea to combine a PS estimator, such as IPTW from 2.15 and a maximum likelihood estimator from a regression model of the outcome, fitted for the estimation of the standardized mean in F.1, the AIPTW estimator for the estimation of m_1 can be derived [46]:

$$\hat{m}_{1;AIPTW} = n^{-1} \sum_{i=1}^{n} \left[\frac{Z_i Y_i}{\hat{e}_i} - \frac{(Z_i - \hat{e}_i)}{\hat{e}_i} m_1(X_i; \hat{b}_1) \right] \tag{2.18}$$

Where $\hat{e}_i := e_i(X_i; \hat{a})$ is the estimated individual PS as defined in 2.2.2. In addition, $E(Y|Z;X) = m_Z(X_i; b_Z)$ is the assumed regression model for the relationship between the vector of covariates and the outcome under treated ($Z = 1$) or untreated ($Z = 0$) conditions. We denote $\hat{b}_1$ the vector of parameter estimates for the outcome model under treatment (i.e., under no treatment, this would be $\hat{b}_0$). Under certain regularity conditions, $\hat{a}$ and $\hat{b}_1$ will converge in probability to some value a^* and b_1^* (irrespective of whether the model is correct or not), with $a^* = a_0$ and $b_1^* = b_{10}$, in the case when these estimators converge to the truth [84].

Now, from the law(s) of large numbers we may replace the $n^{-1} \sum(\cdot)$ by expectations, when n is sufficiently large. In other words, since $\hat{m}_{1;AIPTW}$ is a sample average, it can be shown that it converges in probability (see [81] and Chapter 13 of [84]) to the below expectation:

$$\mathsf{E}\left[\frac{ZY}{e(X;a^*)} - \frac{Z - e(X;a^*)}{e(X;a^*)} m_1(X;b_1^*) \right] \tag{2.19}$$

By the causal assumption of consistency and no interference (or, *SUTVA*, see 2.1.3 and [84]), the observed outcome under treatment may be replaced by the potential outcome we would have observed had we imposed the treatment via the hypothetical intervention we have in mind, i.e.:

$$\frac{ZY}{e(X;a\ )} = \frac{ZY(1)}{e(X;a\ )} = Y(1) + \frac{fZ\quad e(X;a\ gY(1)}{e(X;a\ )} \tag{2.20}$$

Therefore, 2.19 becomes:

$$\begin{aligned} &\mathsf{E}[Y(1)] + \mathsf{E}\ \frac{fZ\quad e(X;a\ )g}{e(X;a\ )}Y(1)\quad \frac{fZ\quad e(X;a\ )g}{e(X;a\ )}m_1(X;b_1) \\ &= \mathsf{E}[Y(1)] + \mathsf{E}\ \underbrace{\frac{fZ\quad e(X;a\ )g}{e(X;a\ )}fY(1)\quad m_1(X;b_1)g}_{augmentation} \end{aligned} \tag{2.21}$$

The first term in equation (2.21), $\mathsf{E}[Y(1)]$, is the expectation of the outcome if everyone in the population were treated. The second term is an augmentation term. If the augmentation term reduces to zero, then equation (2.21) will estimate m_1 (i.e., the expected response if the whole population were to receive treatment). In other words, we need either $a = a_0$ or $b = b_{10}$ (or both), to get a consistent estimate of m_1.

A more intuitive explanation for the derivation of the estimator in (2.18) could be: AIPTW includes two quantities - the first, is the quantity of interest, i.e., the mean outcome if the whole population were to be treated. The second quantity is essentially the product of two bias terms - the bias of the PS model and the bias of the outcome regression model. If the PS model is correct, then bias from this model will be zero, and it will "zero out" the bias from the outcome regression model. Likewise, if the outcome regression model is correctly specified (i.e., zero bias), it will "zero out" the bias from the PS model. In other words, if either the PS or the outcome regression model is correctly defined, the estimator from (2.18) will give an unbiased estimate - the so-called *doubly robust property*.

The proof for statistical consistency of the estimator is not shown herein, however, it can be found, e.g., in [84] or [87] or [46].

Following the same logic, the AIPTW for the estimation of m_0 will be:

$$\hat{m}_{0;AIPTW} = n^{-1}\sum_{i=1}^{n}\frac{(1\quad Z_i)Y_i}{1\quad \hat{e}_i} + \frac{fZ_i\quad \hat{e}_i g}{1\quad \hat{e}_i}m_0(X_i;\hat{b}_0) \tag{2.22}$$

And, to estimate the ATE:

$$\hat{ATE} = \hat{m}_{1;AIPTW} \quad \hat{m}_{0;AIPTW}$$

Lastly, we present the corresponding formulas for each individual for the expected outcome under treated and untreated conditions derived from eq. (2.18), (2.22) below:

Table 2.1 Equations for the expected response under treated ($AIPTW_1$) and untreated ($AIPTW_0$) conditions for each individual in the population

	$AIPTW_1$	$AIPTW_0$
General form	$\frac{Y_{Z=1} \quad Z}{\hat{e}} \quad \frac{\hat{Y}(1)(Z \quad \hat{e})}{\hat{e}}$	$\frac{Y_{Z=0} \quad (1 \quad Z)}{1 \quad \hat{e}} + \frac{\hat{Y}(0)(Z \quad \hat{e})}{1 \quad \hat{e}}$
Among $Z = 1$	$\frac{Y_{Z=1}}{\hat{e}} \quad \frac{\hat{Y}(1)(1 \quad \hat{e})}{\hat{e}}$	$\hat{Y}(0)$
Among $Z = 0$	$\hat{Y}(1)$	$\frac{Y_{Z=0}}{1 \quad \hat{e}} \quad \frac{\hat{Y}(0)\hat{e}}{1 \quad \hat{e}}$

Abbreviations: AIPTW, Augmented Inverse Probability-of-Treatment Weighting; $\hat{e} = P(Z = 1jX)$; Z= treatment; $Y_{Z=0}$ and $Y_{Z=1}$ observed outcome among individuals with $Z = 0$ and $Z = 1$, respectively; $\hat{Y}(0) = \mathsf{E}(YjZ = 0;X)$ = predicted outcome given $Z = 0$; $\hat{Y}(1) = \mathsf{E}(YjZ = 1;X)$ = predicted outcome given $Z = 1$; individual subscript i is suppressed for readability; table inspired by Funk et al., 2011 [46].

Targeted maximum likelihood estimation (TMLE)

Targeted maximum likelihood (TMLE) [88] is an estimation technique that offers an optimal exchange between variance and bias in the estimation of the target parameter (e.g., ATE). TMLEs belong to the class of DR and efficient estimators. The core idea is that for the ATE, TMLE takes initial estimates of $\mathsf{E}[YjZ;X]$ and $Pr(ZjX)$ and afterwards incorporates a substitution *targeting* step which optimises the bias-variance trade-off for the parameter of focus (the ATE in our case). Statistical inference and 95% CIs are based on the efficient influence curve (EIC) theory, however, to delve into more technical details would be outside of the scope of this thesis.

Briefly, the first step of the algorithm is calculating the so-called Q-model (i.e., outcome regression model as a function of treatment and potential confounders), which is essentially the application of standardization, to predict the outcome if the whole population were treated and untreated respectively. This gives an initial estimate $\bar{Q}^0(Z = 1;X)$ and $\bar{Q}^0(Z = 0;X)$, and of the difference between the two, i.e., the ATE. Then, estimation of the PS follows. To counteract the potential residual bias from the association of treatment and potential confounders in our initial estimate $\bar{Q}^0$ from the Q-model, we create the so-called *clever covariates* and estimate the *fluctuation parameter* $\mathsf{e} = (\mathsf{e}_0;\mathsf{e}_1)$. Clever covariates are similar

to the traditional inverse probability-of-treatment weights. A regression model of the outcome Y against the logit of the initial prediction $\bar{Q}^0$ as an offset and the clever covariates (denoted usually as $H(1;X)$, $H(0;X)$) as independent variables is fitted. And this is the key part: if there is actually residual confounding (relationship between the treatment Z and the potential confounders) that was not captured by the initial model, then the clever covariates will account for that by providing an updated estimate (if the PS model is correctly specified). If the initial model was correctly specified, then the fluctuation parameter estimate will be close to zero, because in that case, the PS model would not provide additional information to the initial estimate. There is also the case where the fluctuation parameter estimate might end up close to zero (because the PS might not give additional information to the initial estimate), given although that the initial Q-model is misspecified. SEs may be calculated via the estimation of the variance of the so-called EIC estimate for the ATE based on asymptotic theory. For a smooth introduction into the applied aspect of TMLE, the reader may consult [58].

2.3 Extensions to hierarchical settings

By *hierarchical*, *multilevel*, *clustered* or *nested*, we mean data structures that entail two levels or more, in the sense that lower-level units of analysis that share similar characteristics are grouped together into the same higher-level unit of analysis [66]. Herein, we focus on *two-level* data structures, such as patients nested within hospitals or demographic areas, which is a rather frequent setting in practice[3]. Although hierarchical structures usually entail more than two levels, throughout the dissertation we broadly refer to hierarchical and two-level, interchangeably - more than two levels are not in the scope of the present work. Also, we note that structures different to nested, such as crossed, e.g. when patients may switch to a different hospital, exist too. In such settings, a frequent practice would be to assign the patient to the hospital visited most often; otherwise, methods called *crossed nested random effects* may be required - but, this is not examined herein.

We adopt the term *individual-level*, to describe the lower-level units (e.g., patients) and *cluster-level*, to describe the higher-level units (e.g., hospitals). Previously, we briefly reviewed causal inference methods for single-level observational data (i.e., under independence between individual outcomes). Contrary to that, when patients are nested within the same hospital, they most likely share similar characteristics, introducing a correlation between their observed outcomes. The higher the correlation between individuals of the same cluster, the greater

[3]however, all methods described throughout the thesis are applicable to any context of two-level nested data structures - irrespective of whether it is cancer epidemiology or not.

the heterogeneity between different clusters. Therefore, this correlation between individuals must be accounted for in the analysis to obtain (approximately) unbiased treatment effect and variance estimates.
In the following two subsections we introduce the notation we will be employing in the remainder of the thesis for two-level structures, along with a modification of the identifiability assumptions introduced for the single-level setting in 2.1.3.

2.3.1 Extension of the assumptions

Identifiability assumptions for hierarchical structures and treatment assigned at the individual level

Let us assume an observational cohort study. Consider a sample or population of H clusters, within which, n patients are nested in total. We assume the cluster indicator $h = 1;2;:::;H$ and the cluster-specific size n_h. The total sample size will be $n = \sum_h n_h$. We denote $\mathbf{U}_{hk}$ a vector of unit-level covariates for patient k within cluster h and $\mathbf{V}_h$ a vector of cluster-level covariates for the same cluster.

The identifiability assumptions can therefore be adapted to:

1. Consistency, i.e., for patients that actually received treatment Z of level z their observed outcome is the same as what it would have been had they received treatment level z via the hypothetical intervention we have in mind (or in other words, there are no multiple versions of treatment and observing is the same as imposing):

$$Z_{hk} = z \;) \; Y_{hk} = Y_{hk}(z) \tag{2.23}$$

2. No interference, i.e., the potential outcome $Y_{hk}(z)$ for individual k nested within cluster h, is independent of the treatment level assigned to any other individual j, with $k \neq j$, for $i = 1;2;:::;n_h$.

3. Positivity: i.e., every patient has non zero probability of receiving treatment of level z, for every different level z of treatment Z:

$$Pr((\mathbf{U}_{hk};\mathbf{V}_h) = (\mathbf{u};\mathbf{v})) > 0 \;) \; 0 < Pr(Z_{hk} = zj(\mathbf{U}_{hk};\mathbf{V}_h) = (\mathbf{u};\mathbf{v})) < 1;8z;\mathbf{u};\mathbf{v} \tag{2.24}$$

4. Unconfoundedness or conditional exchangeability or ignorability, or no unmeasured confounding, i.e., the potential outcomes $Y(z)$ are conditionally independent of the

treatment Z given the pretreatment measured covariates $(U;V)$ for every treatment level z:

$$Y_{hk}(z) \;??\; Z_{hk}j(\mathbf{U}_{hk};\mathbf{V}_h);8z \tag{2.25}$$

Where the symbol $??$ indicates statistical independence and the vector $(\mathbf{U}_{hk};\mathbf{V}_h)$ includes the vectors of any potential unit-level and cluster-level confounders, respectively. So, again, this is equivalent to saying that treatment allocation is *at random within different combinations of observed individual- and cluster-level characteristics*.

Alternative formulation of the unconfoundedness assumption within the two-level setting

In practice, important cluster-level confounders remain unmeasured for most observational studies [5]. By conditioning on the cluster, we eliminate any cluster-level confounding, as for a given cluster, any cluster-level variable remains constant. In the existing literature, the previous assumption is sometimes adapted to [18, 96, 48]:

$$Y_{hk}(z) \;??\; Z_{hk}j(\mathbf{U}_{hk};\mathbf{V}_h;a_h);8z \tag{2.26}$$

where, a_h could be either (i) a vector of fixed $H \quad 1$ coefficients indicating cluster membership, *or*, (ii) a vector of random coefficients drawn from a population that compensates for any omitted cluster-level effects. Lastly, $\mathbf{V}_h$ is a vector of any *observed* cluster-level confounders.

Some authors [18], have used the term *latent unconfoundedness*, to describe formulations similar to (2.26). Although the terminology may sound different to (2.25), we still require *all confounders* to be measured and conditioned to the outcome (or the treatment assignment level) to assume treatment levels to be exchangeable between the two treatment groups.

Obviously, if a_h is seen as a constant parameter, specific to the cluster level out of a fixed number of cluster levels, to condition on any cluster-level confounders that are observed in eq. (2.26) is redundant.

If a_h is seen as a random variable, which follows a (prior) distribution that represents any cluster-level effects that are unobserved, to include any measured cluster-level confounders in (2.26) would make sense. However, in that case, the true value of a_h in our sample is *never known*, so, in reality, it cannot be ignored. Yuan and Little have

used the term cluster-specific non-ignorable (CSNI) in the missing data literature, to describe a similar setting [97].

At this point, we must highlight that in the case we base our analysis on models (see [93] on a definition of model-based vs. design-based methods), the required validity of unconfoundedness (as of the rest identifiability assumptions), simply *allows* unbiased estimation. A last (but crucial) requirement, is the validity of the model we apply, and, its corresponding assumptions. We examine a case within which a modeling assumption of random effects models (often overlooked in practice), is violated, in presence of unmeasured cluster-level confounding in Chapter 3.

5. No interference, i.e., the potential individual outcome $Y_{hk}(z)$ for the k^{th} patient nested within the h^{th} cluster is independent of the treatment assigned to any other individual l nested within a cluster m, with $k \neq l$, and $h = m$ or $h \neq m$ (i.e., "the outcomes for each unit are unaffected by the treatment assignments of other units whether within or across clusters", [54]).

Treatment assigned at the cluster level

Although our interest over the following chapters lies within *treatments assigned at the individual level*, for completeness, we present how the identifiability conditions are adjusted for treatments defined at the cluster level.

Identifiability assumptions

Imagine a two-level structure exactly as described in the previous paragraph, although this time, treatment is assigned at the cluster level or, in other words, subjects within the same cluster h are assigned to the same level of treatment $Z_h = z$. An easy way to visualise this design would be to have in mind the equivalent for randomised experiments, which is cluster randomised trials (CRTs), where the randomisation unit is the cluster [38]. In this particular case, the potential outcome of interest might be i. *at the individual*, or, ii. *at the cluster level*, depending on the scientific question of interest [4].

- *Individual-level outcome*

  For this scenario the general identifiability assumptions in 2.1.3 become [89]:

[4] In some particular settings of observational studies, the term 'neighborhood-level interventions' when treatment is assigned at the cluster level is used as well [89].

1. Consistency, i.e., for the clusters that actually received treatment Z of level z the observed individual outcomes are the same as what they would have been had those clusters received treatment level z via the hypothetical intervention we have in mind (again, there are no multiple versions of treatment and observing is the same as imposing):

$$Z_h = z \;) \; Y_{hk} = Y_{hk}(z) \tag{2.27}$$

2. Positivity, i.e. every cluster has non-zero probability of being assigned to treatment of level z, for every different level z of treatment Z:

$$Pr(\mathbf{X} = \mathbf{x}) > 0 \;) \; 0 < Pr(Z = z|\mathbf{X} = \mathbf{x}) < 1; \forall z; \mathbf{x} \tag{2.28}$$

   where $\mathbf{X} = (\mathbf{U}; \mathbf{V})$ is the vector of individual-level and cluster-level covariates.

3. Conditional exchangeability:

$$Y_{hk}(z) \perp\!\!\!\perp Z_h|\mathbf{X}_{hk}; \forall z \tag{2.29}$$

   Where the quantities above are as defined in 2.25.

4. No interference, i.e., the potential individual-level outcome $Y_{hk}(z)$ for the k^{th} patient within the h^{th} cluster is independent of the treatment assigned to any other cluster m, with $m \neq h$. This does not necessarily require that there be no treatment interaction between two subjects within the same cluster, but rather that there be no treatment interaction between individuals in different clusters.

- *Cluster-level outcome*

1. Consistency, i.e., for the clusters that actually received treatment Z of level z their observed outcome is the same as what it would have been had those clusters received treatment level z via the hypothetical intervention we have in mind (again, there are no multiple versions of treatment and observing is the same as imposing):

$$Z_h = z \;) \; Y_h = Y_h(z) \tag{2.30}$$

2. Positivity, i.e., every cluster has non zero probability of being assigned to treatment of level z, for every different level z of treatment Z:

$$Pr(\mathbf{X} = \mathbf{x}) > 0 \;) \; 0 < Pr(Z = z|\mathbf{X} = \mathbf{x}) < 1; \forall z; \mathbf{x} \tag{2.31}$$

where $\mathbf{X} = (\mathbf{U}; \mathbf{V})$ is the vector of individual-level and cluster-level covariates.

3. Conditional exchangeability as in equation 2.25, but, this time defined for the cluster-level outcome:

$$Y_h(z) \perp\!\!\!\perp Z_h | \mathbf{X}_{hk}, \forall z \tag{2.32}$$

4. No interference, i.e., the potential cluster-level outcome $Y_h(z)$ for the h^{th} cluster is independent of the treatment assigned to any other cluster m.

2.3.2 Implications for the statistical analyses

Within the hierarchical (two-level) framework, the key task of the statistical analysis is to account for clustering (i.e. heterogeneity between different clusters, originating from unmeasured cluster-level characteristics plus inherent clustering) [1]. To target a marginal or population average treatment effect, very often a marginal model for the population mean is fitted and Generalized Estimating Equations (GEEs) are used to take into account the correlation of individual outcomes within the same cluster [56].

On the other hand, a different class of models, called *mixed effects models*, targets cluster-specific treatment effects when applied to account for clustering. Nonetheless mixed effects models can still target a marginal treatment effect, just by adding a marginalization step, as we showcase in the next chapter [78, 40, 62]. In comparison to marginal models and GEE methods, mixed effects models allow for direct modeling of the cluster effects. Herein, we focus on these models. A third category that models the cluster indicator is fixed effects models - however, these are unfeasible to apply for moderate to large number of clusters.

In the following chapter, we present how standardization may be applied within the two-level framework when the potential outcomes model is a random effects model.

Chapter 3

Causal inference for hierarchical (two-level) data: Standardization and G-computation

3.1 Introduction

Although literature for handling structured data sets is vast in observational research [66], extensions of methods for hierarchical data in causal inference are relatively new [5, 54, 75, 69, 18, 91]. In this chapter, we focus on how *standardization*, a method that models the outcome mechanism in question (see Section 2.2.1 for the method introduced in the single-level setting), may be extended to the multilevel setting [10, 6]. We primarily target the *marginal or population average treatment effect (p-ATE)* of binary treatments assigned at the individual level on binary or dichotomous outcomes or responses. After distinguishing between *cluster-specific (or conditional)* vs. *population (or marginal) average* treatment effects, we briefly describe three classes of models for the potential outcomes that are frequently applied in epidemiology: the *marginal or single-level model*, the *fixed effects model* and lastly, the *generalized linear mixed (or random) effects model* [67]. We then explain how standardization is performed when we apply the random-effects logistic regression model, which belongs to the class of generalized linear mixed (or random) effects models to model the observed outcomes[1].

[1] under the assumptions of consistency and conditional exchangeability, these correspond to potential outcomes models as well.

Cluster-specific (or conditional) vs *population (or marginal) average* treatment effects

In the existing literature for hierarchical data, the term conditional *vs* marginal effects, traditionally refers to whether we condition within a cluster or not. For instance, at p. 674 of [78], Skrondal and Rabe-Hesketh write *'...Briefly, marginal effects express comparisons of population strata defined by covariate values, whereas conditional effects express comparisons holding the cluster-specific random effects (and covariates) constant...'*. So, generally, in the hierarchical literature, terminology *conditional* vs *marginal* is used with respect to the cluster. With respect to the covariates, these may be assumed to be constant (i.e., we look at effects that are within the same, constant population strata), and in that case we define effects that are not marginal across individuals, but, conditional on some given population strata. However, these may be marginal across clusters (if we average across clusters) or conditional on a given cluster, if we don't. This should be straightforward to understand when looking at non-collapsible effect measures [22], such as odds ratios, for which the conditional odds ratio differs from the marginal odds ratio (with respect to the clusters, covariates or both).

To summarize, the term marginal or conditional effect characterizes the absence or presence of conditioning on: (i) the covariates, when inference is made within the single-level setting or (ii) the clusters, when inference is made within the hierarchical setting. To avoid confusion, herein we distinguish between effects that are:

- conditional on a given cluster and marginal across individuals within that cluster (these will be referred to as cluster-specific average treatment effects or cs-ATE),
- marginal both across clusters and individuals (these will be referred to as marginal or population average treatment effects or p-ATE, i.e., the focal estimand of this dissertation),
- conditional both on a given cluster and the individual characteristics, or,
- marginal across clusters and conditional on individual characteristics.

We primarily target the $p \quad ATE$ across our simulations, as in our settings, the treatment effect is not assumed to be varying between different clusters. The cluster-specific ATE ($cs \quad ATE$) can be also estimated, e.g., when applying clustered estimators (we describe a non-parametric clustered PS weighted estimator [54] in Chapter 4). Regarding interpretation, for the $cs \quad ATE$, imagine that we draw at random a cluster from an assumed population of clusters (or from a fixed number of clusters, each with distinct characteristics). Within that cluster, we calculate the ATE. If we repeat the same procedure, for H clusters with identical cluster characteristics and we average over these, we may get an estimate of the

cluster-specific ATE. Now, if we repeat the previous procedure, for H clusters, drawn from the target population of clusters, and then, average over the $cs \quad ATEs$, we may get the marginal (averaged over both clusters and individuals) ATE or $p \quad ATE$ estimate. If the cluster size varies across clusters (called an unbalanced design in multilevel data literature), then, we may account for that in the final $p \quad ATE$ estimate.

3.2 Standardization for two-level structures

We now briefly describe how standardization (Chapter 2) is extended to two-level structures. This is primarily based on [66], [67], [78] and Chapter 13 of [43].

3.2.1 Estimation of Average Treatment Effect (ATE)

Before we proceed, let us provide some motivation: recall that our target is the $p \quad ATE$, hence, a natural extension for standardization in the two-level setting may be motivated by expression (2.8) for the $p \quad ATE$. To simplify, let us focus on the population mean under treatment, i.e.:

$$E[Y(1)] = E[E(YjX;Z = 1)] \tag{3.1}$$

Intuitively, in presence of clustering, one should opt for modeling tactics to model $E(YjX;Z = 1)$, while accounting for clustering[2]. To get the outer expectation in the right-hand side of (3.1), we should average over the predictions under treatment, across individuals to obtain the $p \quad ATE$.

Now, focusing on the particular steps of standardization (Chapter 2):

i. Model the outcome as a function of both the treatment and the postulated confounders.

ii. Predict the outcome for each individual as if everyone were treated (respectively, untreated).

iii. Average these predictions for the treated (respectively for the untreated) over the sample [3] of individuals.

iv. Contrast these averages according to the estimand and effect measure of interest (e.g., risk difference when targeting the $p \quad ATE$).

[2] essentially, in our settings we model $logit[E(YjX;Z = 1)]$.

[3] where the sample drawn, is assumed *representative* of our target population.

In other words for step i, we must opt for a model that accounts for the correlation between individuals of the same cluster. For step ii, depending on the model fitted, we must decide on how to predict the individual probabilities of the outcome[4] (under treatment or control, respectively) - which in turn, is decided by the target estimand of inference (e.g., marginal or population average treatment effect - p-ATE). Hereafter, we elaborate on these two points.

Outcome models for binary responses

The modeling choice for the outcomes depends mainly on whether we have (i) measured all possible cluster-level confounders and (ii) the number of different clusters is that high that requires too many different regression parameters to estimate from our respective regression model. We present the different modeling options below:

1. A marginal or single-level outcome model [5]:

$$logit\,fPr(Y_{hk} = 1 j Z_{hk} = z; \mathbf{X}_{hk} = \mathbf{x})g = h_0 + Z_{hk}g + \mathbf{X}_{hk}b \tag{3.2}$$

   Where h_0 is the fixed intercept and g is the constant treatment effect; b is the vector of the regression coefficients, and, lastly, $\mathbf{X}_{hk} = (\mathbf{U}_{hk}; \mathbf{V}_h)$, the vector of requisite measured individual- and cluster-level confounders, respectively. We define as $logit(x) = \frac{x}{1\ x}$. Here, it is assumed that the effect of the cluster on the outcome is only through the *observed* cluster-level covariates. Correlation between units is seen as a nuisance feature and only addressed by the variance estimation for the p-ATE, e.g, by applying GEEs (see [56, 77] and subsection 2.3.2).

2. A fixed-effects outcome model:

$$logit\,fPr(Y_{hk} = 1 j k_h; Z_{hk}; \mathbf{U}_{hk})g = k_h + Z_{hk}g + \mathbf{U}_{hk}b \tag{3.3}$$

   Where k_h is the cluster-specific main effect that not only absorbs all the between-cluster variability, but also eliminates any cluster-level confounding, since we condition on a given cluster, h. This is an attractive choice for cases where we have (i) few clusters and (ii) each cluster is seen as a distinct entity, which would forbid exchangeability of

[4] under the assumptions of consistency and conditional exchangeability, these correspond to potential outcomes as well.

[5] when referring to marginal, we mean *marginal with respect to the cluster, i.e., across clusters*, contrast to *conditional on a given cluster, which is the case for the next two models; nonetheless, this model continues to be conditional on the covariates*.

cluster id labels - e.g., clusters being a set number of distinct countries. In this case however, it is impossible to estimate any specific cluster-level covariate effects. If clusters are too many (which might be the case for studies that aim to reduce inequalities in cancer outcomes), the high number of regression parameters to estimate in the model above creates unstable regression coefficient estimates - a phenomenon that falls under the headline of the so-called Neyman-Scott incidental parameter problem [61]; in that case, clusters may be alternatively seen as a sample drawn from a population with similar characteristics, which leads to the next modeling alternative.

3. A random-effects logistic regression outcome model [6]:

$$logit\{Pr(Y_{hk} = 1|h_h; Z_{hk}; \mathbf{X}_{hk})\} = r^{0}_{hk}h_h + Z_{hk}g + \mathbf{U}_{hk}b + \mathbf{V}_h d \tag{3.4}$$

Where the vector for the random effects h_h conventionally follows a joint multivariate normal distribution: $h_h \quad MVN(\mathbf{0}; S_h)$, with S_h the variance-covariance matrix for the random effects. The random effects h_h are assumed independent and identically distributed across clusters h and independent of the covariates; $\mathbf{V}_h$ is the vector of any observed cluster-level covariates.

In the simplest setting, we may define a random-intercept logistic regression [7] outcome model. In that case, h_h is one-dimensional. In the multidimensional case, we may have random slopes for some of the included covariates or treatment, if their effects are assumed to differ between clusters. For example, if we assume a random slope for one of the individual-level covariates, then we have a random intercept and slope model, where h_h is a two-dimensional, column-vector. In the next chapter, we assume settings with random intercept and slope models in some of our investigated simulation scenarios.

For all the above models, statistical interactions may be added between covariates and other covariates or treatment (e.g., if effect modification is assumed).

[6] A random-effects logistic outcome model falls under the headline of generalized linear mixed (or random) effects models, which are random-effects models applied to model responses that are non-linear (such as binary); the latter defines the choice of the link function, $g(\cdot)$, which could be, e.g., the $logit(\cdot)$, $probit(\cdot)$, $clog \quad log(\cdot)$, - herein, we focus on the $logit(\cdot)$ link [67]: $g\{Pr(Y_{hk} = 1|h_h; Z_{hk}; \mathbf{X}_{hk})\} = r^{0}_{hk}h_h + Z_{hk}g + \mathbf{U}_{hk}b + \mathbf{V}_h d$.

[7] An alternative formulation, often used in econometrics and psychometrics, is the latent-response variable formulation; however, generalized linear (mixed effects) models are more common in statistics and biostatistics [67]; therefore, we retain the formulation introduced in 3.4 throughout this dissertation.

Assumptions of the random effects model

Hereafter, we focus on random (or mixed) effects models to model the outcome. Therefore, we must present the modeling assumptions that need to be met, in order to obtain consistent regression coefficient estimates from a logistic mixed effects model. As Rabe-Hesketh and Skrondal state [67], the random part of the model needs to be correctly specified, in addition to the fixed part (in comparison to continuous outcomes, where only correct specification of the fixed part should suffice); So, to get consistent estimates of the regression coefficients of a random effects logistic model, the below should be met:

i. Correctly specified linear predictor (i.e., the right-hand side of the equality in (3.4)),

ii. Correct link function (in our settings, this is assumed to be a logit, however other choices like probit or clog-log are possible),

iii. Correct specification of covariates having random coefficients (in this chapter, we do not assume covariate effects to differ between different clusters, hence our models include only a random intercept; in section 4.4 the effect of one of the individual-level covariates varies across clusters, therefore we add a random slope for that covariate),

iv. Individual responses to be independent conditional on the covariates and the random effects,

v. The random effects and any included covariates are independent (*for causal inferences*)[8] - broadly referred to as *exogeneity*,

vi. Random effects are normally distributed.

When a cluster-level confounder, which, by definition affects both the treatment assignment and the outcome mechanism, is not included in the model, then, assumption (v.) no longer holds. The extent this violation might affect the $p \quad ATE$ estimate, is examined in a short communication paper draft, with intent to be submitted for consideration for publication. This is presented in Section 3.3.

[8] the key assumption to causally interpret a coefficient estimate obtained from a linear regression model is referred to as *strict exogeneity* i.e., it is required that $E[e_i^c | T_i^c] = 0$, with e_i^c the individual residuals and T_i^c the treatment values. Strict exogeneity implies that there is no correlation between unmeasured effects (if we see the residuals e_i as representing the *combined effects of all unmeasured variables*), and the included covariates in the model [66]. Over the next chapters, we require the latter, referring to it broadly as *exogeneity*, and any violation of it as *endogeneity*.

Methods to predict the individual probabilities of response for random-effects models

Unless otherwise stated, this paragraph was based on [40] and [62]

A. Predict the *population average (PA)* or *marginal* probabilities of response for each individual

Let us introduce some notation. Let $\hat{b}^{cs}$ be the cluster-specific coefficient estimates from a logistic random-effects model for the outcome. Let also, $\hat{T}$ be the Cholesky factorization matrix[9] estimate of the variance-covariance matrix for the random effects, $\hat{\mathrm{S}}_h$. For the simplest case of a random intercept model, we denote as $\hat{s}^2_h$, the variance estimate of the random intercepts. In addition, let $\hat{b}^{pa}$, be the population averaged coefficient estimates for the equivalent marginal model. We may calculate $\hat{p}^{pa}_{hk}(1)$ and $\hat{p}^{pa}_{hk}(0)$, i.e., the estimates of the *population strata specific but marginal on the clusters* probabilities of the potential outcome under treated or untreated, correspondingly - as averaging over individuals and contrasting between treated and untreated will ultimately provide the *marginal or population average ATE estimate ($p\quad ATE$)*, which is our target of inference and marginal over both clusters and individuals (recall: steps iii. and iv. in 3.2.1):

$$p\quad \backslash ATE = \frac{1}{N}\sum_{h=1}^{H}\sum_{k=1}^{n_h}(\hat{p}^{pa}_{hk}(1)\quad \hat{p}^{pa}_{hk}(0)) \tag{3.5}$$

For the simplest case of a random intercept logistic model, there is an approximation formula to obtain the PA coefficient estimates, $\hat{b}^{pa}$ [3]:

$$\hat{b}^{pa}\quad \hat{b}^{cs} = \frac{\sqrt{\hat{s}^2_h + p^2/3}}{p^2/3} \tag{3.6}$$

where $p^2/3$ is the variance of the standard logistic distribution. Others suggest multiplying the logistic variance by the factor 15/16, a result obtained from a cumulative Gaussian approximation to the logistic function [98]. After having obtained the population average regression coefficients, $\hat{b}^{pa}$, one may apply the $expit(x)$ function, where $expit(x) = \frac{exp(x)}{1+exp(x)}$, to transit to the probability scale.

[9] i.e., for a matrix of dimension $r\quad r$, S_h, $9T$, such that S_h can be written as the product of T and its transpose, T^t: $\mathrm{S}_h = T\quad T^t$; in other words, T is the matrix equivalent of the square root; this formulation helps the calculations applied in the following - e.g., see formulation in (3.8).

For multiple random effects models, such as a random intercept and random slope model, there is no approximation formula similar to the above. Instead, one has to solve the below defined integral (proposed by Griswold and Zeger and others [35]):

$$\hat{p}_{hk}^{pa}(z) = \int_{q \in Q} expit(x_{hk}^t \hat{b}^{cs} + z_{hk}^t \hat{T} q_h) dF(q), \text{ for } z = 0; 1 \tag{3.7}$$

Where T the Cholesky factorization of the random effects variance-covariance matrix as before, q_h the vector of the standardized random effects, i.e., these jointly follow a standardized multivariate normal distribution[10], F the cumulative distribution of q_h and Q the support space of the distribution of the vector of the random effects. The differential $dF(q)$ can be thought of as a shorthand for $[F_q(q_i) \quad F_q(q_{i\ 1})]$[11].

Since it is impossible to solve the previous integral analytically (see, for instance [86]), one approach would be to approximate it via a sum, by applying Gauss–Hermite quadrature. Some alternatives are mentioned in [78], Section 2: *"...typically by adaptive quadrature (e.g. Pinheiro and Bates (1995) and Rabe-Hesketh et al. (2005)) or by Monte Carlo integration (e.g. McCulloch (1997)). Alternatives to maximum likelihood that do not require integration include penalized quasi-likelihood (e.g. Breslow and Clayton (1993)) and Markov chain Monte Carlo sampling (e.g. Clayton (1996))..."*. Another alternative mentioned in the supplementary material of [39] is Romberg integration. Methods like, Markov chain Monte Carlo (MCMC) are computationally demanding. We now present Gauss-Hermite quadrature (also applied in the *R* package *lme4*), which is widely applied in practice:

$$\hat{p}_{hk}^{pa}(z) \quad \sum_{q=1}^{Q^r} expit(x_{hk}^t \hat{b}^{cs} + z_{hk}^t \hat{T} B_q) A(B_q), \text{ for } z = 0; 1 \tag{3.8}$$

Where, r is the number of random effects (e.g., for a random intercept and random slope model, $r = 2$), Q the number of quadrature points, B_q, $A(B_q)$ the optimal quadrature points and weights for the standard normal univariate density [83].
For r random effects, we have a multivariate standard normal density. In our case, when $r = 2$: $B_q^{\emptyset} = (B_{q_1}; B_{q_2})$, with its associated (scalar) weight, given by the product of the corresponding univariate weights:

[10] the product Tq_h gives us the original form of the random effects, h_h; this transformation simplifies the numeric calculation of the integral via the sum in 3.8.

[11] this is a more general formulation of the so-called Riemann integral, known as Riemann-Stieltjes integral; e.g., see https://www.statlect.com/fundamentals-of-probability/expected-value.

$$A(B_q) = \bigotimes_{h=1}^{2} A(B_{q_h}) = A(B_{q_1})A(B_{q_2}) \tag{3.9}$$

A second approach to approximate the integral in (3.7) is to apply Monte Carlo integration [86], which is the method we applied in our simulations settings presented in sections 3.3 and 4.4 [12]. The main idea is that (3.7) can be seen as the mean value $E_{q_h}[\Upsilon(x_{hk}^t \hat{b}^{cs} + z_{hk}^t \hat{T} q_h) dF(q)]$ averaged over the distribution of q_h. In that manner, one may estimate this mean value by the sample mean obtained from n simulated samples, with n sufficiently large. This is characterized as a "stochastic approach", as the resulting value depends on the number of simulations, with the variability stemming from the sampling process decreasing as the number of simulations increases. The utility of this method over Gauss–Hermite quadrature becomes obvious as the dimension of the integral increases.

Formula 3.7 targets the expected probability of response under treatment level z, for a unit randomly drawn from the overall population, in a new cluster - since we average over the assumed *prior* distribution of the random effects.

B. Predict the probabilities of response for each individual conditional on the *Empirical Bayes Estimates (EBE) for the random effects*

Let us denote $\hat{p}_{hk}^{ebe}(z)$, the estimated (or predicted) probability of response for individual k nested in cluster h that is conditional on the random effects vector under treatment level z. Since the specific values for the random effects vector are not known, one may plug-in their estimates for each specific cluster. One may obtain the so-called EBEs for the random effects and plug-in these:

$$\hat{p}_{hk}^{ebe}(z) = expit(x_{hk}^t \hat{b}^{cs} + r_{hk}^t \hat{d}_h^{ebe}), \text{ for } z = 0; 1 \tag{3.10}$$

An elaborate explanation of how the Empirical Bayes Estimates are obtained is provided in Section 4 of [78]. Briefly, the name *Empirical Bayes* originates from the fact that the method applied to get the estimates $\hat{d}_h^{ebe}$ is semi-Bayesian: although the estimates of the random effects come from their estimated posterior means, the way the posterior distribution is calculated is not entirely Bayesian. As they write in [78]: *'...In a fully Bayesian approach, prior distributions would be specified for the model parameters, and the posterior distribution of the random effects would be marginal*

[12] In our simulations, we applied the *R* package *GLMMadaptive*, which makes use of Monte Carlo integration [68]

with respect to these parameters'. Contrary to this, any unknown parameters are simply replaced by their estimates from the data; in particular, the vector of the outcome model parameters b, is replaced by the respective maximum likelihood estimates (MLE) $\hat{b}$. To obtain the posterior distribution w of the random effects, we need two sources of information: (i) prior knowledge for the clusters before looking at the data (corresponding to the postulated prior distribution of the random effects), and, (ii) information provided by the data, i.e., the outcome vector $\mathbf{y}_h$, the covariate matrices $\mathbf{X}_h$ and $\mathbf{R}_h$, with the latter being the matrix of covariates with random effects, all for cluster h:

$$w(d_h|\mathbf{y}_h;\mathbf{X}_h;\mathbf{R}_h;\hat{b}) = \frac{f(d_h;\hat{\mathsf{S}}_d)f(\mathbf{y}_h|d_h;\mathbf{X}_h;\mathbf{R}_h;\hat{b}^f)}{g(y_h|\mathbf{X}_h;\mathbf{R}_h;\hat{b})} \tag{3.11}$$

Where, f, the assumed prior distribution of the random effects (e.g., a multivariate normal in our settings), f the distribution of the response, given the random effects, covariate values and the model parameter MLEs. Lastly, g is the likelihood of the responses, given the covariate matrices and the estimated parameters $\hat{b}$.

Once we have obtained the EBE for each cluster, we may calculate each $\hat{p}_{hk}^{ebe}(z)$ for every treatment level. Then, we may estimate the ATE within each cluster:

$$\widehat{ATE}_h = \frac{1}{n_h}\sum_{k=1}^{n_h}\hat{p}_{hk}^{ebe}(1) \quad \frac{1}{n_h}\sum_{k=1}^{n_h}\hat{p}_{hk}^{ebe}(0) \tag{3.12}$$

Note that if we assume the treatment effect to vary across different clusters (i.e., a random slope for the treatment), then we expect the values of (3.12) to differ across different clusters, in which case, cluster-specific effects would be an important target of inference, on top of an average *ATE* estimate across clusters (i.e., a $p \quad ATE$).

Lastly, we may average the cluster-specific ATEs over the total number of clusters to estimate the marginal or population averaged ATE. In particular, we may write:

$$\widehat{ATE} = \frac{1}{H}\sum_{h=1}^{H}\widehat{ATE}_h \tag{3.13}$$

If we have a balanced design, i.e., each cluster consists of the same number of individuals or units or clusters, $n_h = n$:

$$\widehat{ATE} = \frac{1}{N}\sum_{h=1}^{H}\sum_{k=1}^{n_h}(\hat{p}_{hk}^{ebe}(1) \quad \hat{p}_{hk}^{ebe}(0)) \tag{3.14}$$

Formula 3.10 targets the expected probability of response under treatment level z, for a unit randomly drawn from within (given) a hypothetical cluster - that is, since we condition on the empirical Bayes estimates for the random effects; we note that an also widely applied approach is to consider the median (i.e., a random intercept equal to zero) in random intercept models [78], which also targets a hypothetical cluster; we present that in the following paragraph.

C. Predict the probabilities of response for each individual *conditional on the value of zero for the random effects*

Equality (3.10) becomes:

$$\hat{p}_{hk}^{\hat{d}_h=0}(z) = expit(x_{hk}^t \hat{b}^{cs}), \text{ for } z = 0;1 \tag{3.15}$$

Formula (3.15) targets the expected (conditional on covariates) probability of response under treatment level z, for a unit in a hypothetical cluster [78].

We note that, after averaging over individuals, all the aforementioned prediction approaches, A, B, and C, ultimately estimate the $p \quad ATE$.

Other practices are to predict the probabilities of response for each individual by marginalizing over *the posterior distribution for the random effects* [78]. This targets the marginal (across clusters) prediction of response under treatment level z, for a new unit in an existing cluster.

3.2.2 Estimation of variance

To estimate the variance of the p-ATE estimators, one may apply the non-parametric bootstrap, where we may sample with replacement the independent units, i.e., the clusters, and apply the corresponding method to obtain the *marginal or population averaged ATE* within each bootstrap sample. More specifically:

The Bootstrap is a general simulation method that may be applied to obtain an approximate estimate for the variance of a parameter in cases where a closed-form estimator is not feasible or straightforward to generate. A brief intuition of the method in the single- level setting is provided in Chapter 5.

In our two-level settings we assume correlated units (i.e., patients) to be nested within independent clusters (e.g., hospitals). Since the bootstrap principle requires observations to be independent (see Chapter 5), an approach to perform the Bootstrap is to

sample with replacement the clusters (i.e., the independent units). We may iterate this procedure up to B iterations, which gives us the number of bootstrap samples. The number of bootstrap samples is advised to be relatively high [17]. For each bootstrap sample, we may perform the standardization step under treatment and under no treatment, and then take the contrast between the two to obtain the $\widehat{ATE}$ via methods A, B or C.

3.3 Research Paper I

Below, we present a draft of a short communication about to be submitted that investigates the estimation of the $p \quad ATE$ with mixed effects models, under exclusion of a cluster-level confounder - which violates the independence assumption between the random effects and any covariates included in the mixed effects model. Corresponding code to generate and analyse the simulated data sets of the simulations performed in the paper, can be found at the corresponding GitHub repository sims-gcomp. A corresponding appendix of the paper draft can be found in Appendix A.

In the following chapter, we present how *IPTW* an *AIPTW* may be applied within the two-level framework, when the PS model is a random effects model.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	lsh2005314	**Title**	Ms
First Name(s)	Andriana		
Surname/Family Name	Kostouraki		
Thesis Title	Comparison of causal inference methods for observational data with a hierarchical structure		
Primary Supervisor	Aurelien Belot		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Not decided yet
Please list the paper's authors in the intended authorship order:	Andriana Kostouraki, Clémence Leyrat, Aurélien Belot
Stage of publication	**Not yet submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I was the lead author of this paper. The concept was developed in collaboration with the Co-authors. I drafted the manuscript. Co-authors contributed to the simulation design and analysis, critically revised the manuscript and offered helpful insight and feedback on the manuscript.

SECTION E

Student Signature	[redacted]
Date	16/07/2024

Supervisor Signature	[redacted]
Date	16/07/2024

Received <day> <Month>, <year>; | Revised <day> <Month>, <year>; | Accepted <day> <Month>, <year>

DOI: xxx/xxxx

SHORT COMMUNICATION

A note on the estimation of marginal causal e□ects with mixed e□ects models

Andriana Kostouraki*[1] | Clémence Leyrat[2] | Aurélien Belot[1]

[1]Inequalities in Cancer Outcomes Network, Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT

[2]Department of Medical Statistics, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT

Correspondence
*Andriana Kostouraki, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT. Email: andriana.kostouraki@lshtm.ac.uk

Abstract

When analysing the causal e□ect of an individual-level treatment in hierarchical observational data, cluster-level confounders may remain unmeasured, violating the exchangeability assumption - a.k.a unconfoundedness - which is required for unbiased estimation. In this context, unconfoundedness may be seen as “latent”, as within cluster, treatment groups are exchangeable given *all* individual-level confounders. Once having obtained unbiased estimates for the cluster-specific average treatment e□ects (cs-ATE), one could average over the distribution of clusters to get an unbiased population ATE (p-ATE) estimate under certain assumptions. In the presence of clustering, mixed e□ects models are often used to account for correlation between individual outcomes within clusters. Clusters are directly modelled via random e□ects. Nonetheless, independence between the random e□ects and the included covariates must hold to secure consistent coe□cient estimates. We examine the impact of unmeasured cluster-level confounding when estimating the p-ATE. We investigate scenarios where we apply G-computation (standardization) combined with mixed e□ects. A simulation study explores scenarios of an unobserved cluster-level confounder. We consider a binary treatment assigned at the individual level and a binary outcome. We analyse the data using outcome models with random e□ects to account for clustering. Although random e□ects models provide biased cluster-specific and conditional on population strata e□ect estimates when omitting cluster-level confounders, our simulations under the null indicated this bias decreased when the target was the p-ATE. The same pattern arose in presence of correlation between the omitted cluster-level confounder and an individual-level covariate.

KEYWORDS:
Causal Inference, Cluster-level confounding; G-computation; Multilevel models; Observational Studies, Unconfoundedness

1 | INTRODUCTION

Unlike randomised controlled trials, observational studies may su□er from systematic di□erences in the distribution of patient characteristics between treatment groups. To address measured confounding, causal inference methods are mostly developed in the single-level setting, i.e. when the treatment, outcomes and covariates are measured at the same unit level and observations are independent.

However, clustering often occurs naturally while being pervasive in observational data - e.g., when patients are nested in hospitals[1,2,3] - and individual outcomes within the same cluster are correlated. In the presence of clustering, two causal estimands can be targeted: the cluster-specific e□ect and the population-average e□ect (both being either marginal or conditional on covariates).

To account for clustering, one may apply mixed-e□ects models, where random e□ects capture the between-cluster heterogeneity, originating from unmeasured cluster characteristics. By definition, the random e□ects are assumed independent from the covariates included in the model (exogeneity)[4]. Mixed-e□ects models allow a direct estimation of cluster-specific e□ects (marginal or conditional on covariates), but population-average e□ects can still be obtained by adding a marginalization step, which is rarely detailed in the multilevel causal inference literature.

Cluster-level confounders often remain unobserved[5]. Using the potential outcomes framework, we discuss: (i) the impact of ignoring clustering in the analysis of observational data and (ii) whether mixed-e□ects models can account for unmeasured cluster-level confounding when targeting population-average and cluster-specific e□ects of an individual-level intervention. After introducing the notations and estimands, we present two approaches to estimate the *population-average treatment e ect* (p-ATE). In simulated scenarios with unmeasured cluster-level confounding, we evaluate the performances of di□erent estimators, and discuss the practical implications of these findings.

2 | CAUSAL FRAMEWORK FOR TWO-LEVEL STRUCTURES: NOTATIONS, ESTIMANDS AND ASSUMPTIONS

Assume a cohort study of □ patients nested within □ = 1, 2, ..., □ hospitals of size □□. We denote $U_{□□}$ and $V_{□}$ the vectors of individual- and cluster-level covariates for patient □ within cluster □, respectively. $□_{□□}$ is the individually assigned binary treatment for patient □ within cluster □, while, $□_{□□}(1)$ and $□_{□□}(0)$ are the individual potential outcomes as if they were treated or untreated, respectively. The population-averaged treatment e□ect is:

$$\text{p-ATE} = □[□_{□□}(1) − □_{□□}(0)]. \tag{1}$$

This estimand represents the average e□ect in the target population.

In some settings, the cluster-specific ATE is of additional interest:

$$\text{cs-ATE} = □[□_{□□}(1) − □_{□□}(0)|□], \tag{2}$$

which represents the average treatment e□ect between two patients drawn at random from the same cluster, i.e., the average treatment e□ect conditionally on a cluster.

These causal e□ects can be identified from the data if the standard causal assumptions of consistency[6], positivity[7], conditional exchangeability[8] hold at both the individual and cluster levels (more details in Appendix). According to some, causal consistency involves both the assumptions of (i) treatment variation irrelevance, and (ii) no interference[9]. Herein, assuming that the rest assumptions hold, we focus on conditional exchangeability or unconfoundedness:

$$□_{□□}(□) ⫫ □_{□□}|(U_{□□}, V_{□}), □□, \tag{3}$$

where ⫫ indicates statistical independence. This means that treatment is allocated *at random within each combination of observed individual- and cluster-level characteristics*.

A full specification of cluster-level characteristics for a given cluster is equivalent to conditioning on that cluster[5]. Hence, when some cluster-level confounders are unavailable, we may require *latent unconfoundedness* instead[10]:

$$\square_{\square\square}(\square) \perp\!\!\!\perp \square_{\square\square} | (U_{\square\square}, V_{\square}, \square_{\square}), \square\square \tag{4}$$

where, $\square_{\square}$ is cluster-specific and it could be (□) a vector of fixed $\square - 1$ coe□cients (indicating cluster membership), *or*, (□□) a random variable drawn from a given distribution. If a cluster-level confounder is omitted and we do not condition on the cluster, assumption (4) is violated leading to bias when estimating the p-ATE. We examine the impact of ignoring clustering and whether mixed e□ects models help eliminating or reducing this bias.

3 | ESTIMATION OF THE P-ATE FOR TWO-LEVEL STRUCTURES

G-computation[11] is a popular method to estimate marginal e□ects from conditional regression models. While g-computation is well-described for generalised linear models, its implementation for generalised linear mixed-e□ects models is less known within the field of epidemiology. A marginalisation over the covariate distribution would only lead to estimating the marginal cs-ATE. A further marginalisation over the cluster distribution would lead to estimating the p-ATE. We now present two approaches to perform this marginalisation to estimate the p-ATE.

A first approach is to integrate over the prior distribution of the random intercepts[12]. This targets the predicted probability of response as if one were treated, $\hat{\square}^{\square\square}_{\square\square}(1)$ (or respectively untreated, $\hat{\square}^{\square\square}_{\square\square}(0)$) for *a unit drawn from a new cluster, randomly selected among the population of clusters*[12,13]:

$$\hat{\square}^{\square\square}_{\square\square}(\square) = \int_{\square\square\square} \square\square\square\square(\square'_{\square\square}\hat{\square}^{\square\square} + \square_{0\square})\square\square(\square), \square = 0, 1, \tag{5}$$

where $\square\square\square\square(\square) = \frac{\square\square\square(\square)}{1+\square\square\square(\square)}$, $\hat{\square}^{\square\square}$ the cluster-specific coe□cient estimates from a logistic mixed-e□ects model and $\square_{0\square}$ the random intercepts, with $\square_{0\square} \square \square(0, \square_{\square})$ - where $\square_{\square}$ the variance of the random intercepts. $\square$ is the support space of the random intercepts distribution, $\square_{\square\square}$ the vector of covariate values (including treatment $\square$), and, $\square$ the cumulative density function of the random e□ects.

The integral in eq. (5) cannot be calculated analytically, but can be approximated numerically[12,14]. The R package GLM-Madaptive[15], used to analyse the simulated datasets in section 4, applies Monte-Carlo integration.

A second approach is to predict the individual probability of response, based on the respective cluster the individual belongs to[16,17]. Since the true value for the cluster is unknown, one may plug in the empirical Bayes estimates (EBE) for the random e□ects:

$$\hat{\square}^{\square\square\square}_{\square\square}(\square) = \square\square\square\square(\square'_{\square\square}\hat{\square}^{\square\square} + \hat{\square}^{\square\square\square}_{\square 0}), \square = 0, 1, \tag{6}$$

where $\hat{\square}^{\square\square\square}_{\square 0}$ the EBE of the random intercepts. This targets the individual probability of response for *a new unit from an existing cluster*.

Once having estimated the individual probabilities of response under treatment and no treatment from (5) or (6), we may average over individuals for each treatment-level and take the contrast between the two to estimate the p-ATE. In the following, the performance of the two approaches is empirically compared.

4 | EMPIRICAL EVALUATION OF MIXED EFFECTS MODELS UNDER OMISSION OF A CLUSTER-LEVEL CONFOUNDER

4.1 | Methods

Our simulation study is fully described in the Appendix, following current guidelines ([18]). Briefly, treatment allocation and potential outcomes were generated from random intercept models under two data-generating mechanisms, where a cluster-level confounder could be correlated or not with one of the two individual-level confounders.

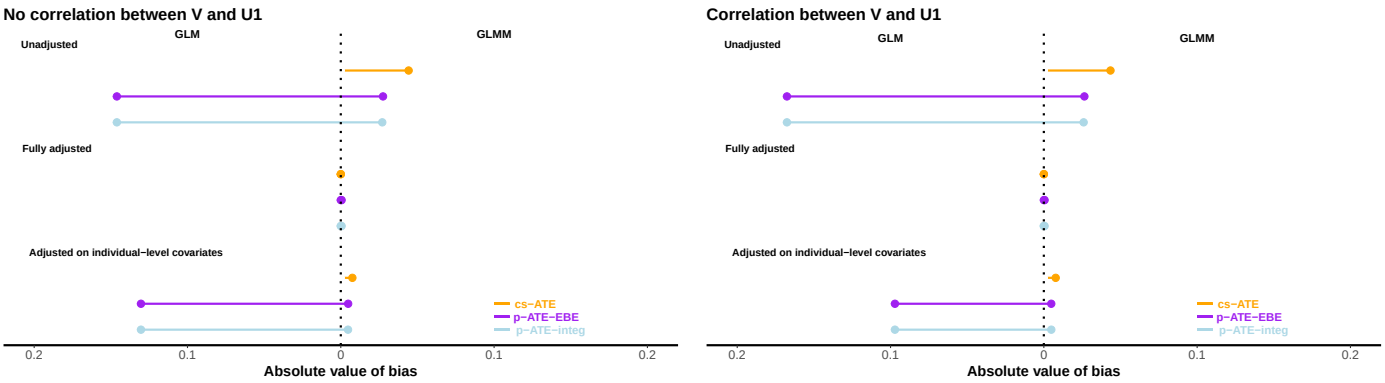


FIGURE 1 Absolute value of bias across DGMs (DGM 1 and 2 from left- to right-hand side; marginal event rate of 45%) for analysis methods: () unadjusted, () fully adjusted, () adjusted on individual-level covariates, combined with either . GLM (Generalised Linear Model) or, . GLMM (Generalised Linear Mixed e□ects Model); DGM: Data Generating Mechanism; $□_{□□□} = 1,000$ simulations of a balanced design of 100 clusters of 50 patients; all DGMs included a random intercept; note: p-ATE EBE and p-ATE integ coincide for the GLM analyses; correlation = 0.5 between $□_1$ and .

We illustrate how: () the p-ATE and the cs-ATE estimates are impacted under omission of a cluster-level confounder, () the two approaches to estimate the p-ATE comparatively perform and () ignoring clustering altogether a□ects the p-ATE estimate. We considered unadjusted, adjusted for the individual-level covariates and fully adjusted analyses, and we focused on bias and coverage rate.

4.2 | Results

The fully adjusted GLMM corresponds to the data generation model. It is therefore the *benchmark model*, and, unbiased for all estimands.

Ignoring clustering

As expected, both the unadjusted GLM and GLMM provided biased estimates due to introduction of confounding bias. However, the bias of the unadjusted GLMM was lower as the random intercept had absorbed some of the unexplained between-cluster heterogeneity attributed to the unobserved confounder . The fully adjusted GLM is unbiased as it captures all the individual and cluster-level confounding, but we know from theory that standard errors would be incorrect, clustering being ignored.

EBE vs. integration

Bias of the p-ATE was nearly identical in all scenarios when plugging-in the Empirical Bayes estimates of the random e□ects (ATE-ebe) or integrating over the prior distribution of the random intercepts (ATE-integ).

Omitting a cluster-level confounder

When omitting V, the bias increased for both the cs-ATE and p-ATE, in comparison to models including V. When was correlated with $□_1$, the cluster-specific coe□cient estimate of $□_1$ was biased (see appendix, Table 1 *vs* Table 2) as it violates the random-e□ects exogeneity assumption. This could impact both the p-ATE and the cs-ATE - as the cluster-specific coe□cients are applied to calculate both the p-ATE and the cs-ATE in (5) and (6).

5 | CONCLUSION

We observed that, when targeting the p-ATE and the cs-ATE (both marginal on covariates), the bias decreased when using a mixed e□ects model (under the assumption of no treatment e□ect) compared to a model ignoring clustering. We have considered two di□erent approaches for G-computation when applying the random intercept models: () integrating over the prior distribution of the random intercept[12], or, () plugging-in the EBEs of the random intercept to predict individual probabilities

of response. Bias from the two approaches was nearly identical across scenarios. However the two approaches theoretically target di□erent probabilities: the former targets the probability of response (conditionally on treatment) in a *new* cluster, while the latter targets the probability of response (conditionally on treatment) in an *existing* cluster. For the cs-ATE we predicted the individual probability of response, conditional on the treatment level on a given cluster.

If a cluster-level variable is omitted from the analysis, the random e□ects should represent this omitted variable, which however might be associated with treatment and so being a confounder. Nevertheless, in theory, random e□ects cannot rectify for any omitted cluster-level confounders. Solutions for exogeneity violations when targeting cluster-specific coe□cients were discussed elsewhere[3]. We instead, focused on the marginal or population ATE and the cluster-specific (but marginal on covariates) ATE.

To conclude, in most observational studies, there is clustering that is rarely accounted for, and we have shown that this can lead to substantial bias. Including a random-e□ect (whatever the estimand) reduced greatly this bias in our scenarios. This result coincides with previous studies (e.g., [19]).

6 | ACKNOWLEDGEMENTS

AK is supported by the Cancer Research UK Doctoral Studentship in Inequalities in Cancer Outcomes (CR UK Programme C7923/A30945). CL is supported by the UK Medical Research Council (Skills Development Fellowship MR/T032448/1) and by the European Union's Horizon 2020 research and innovation programme under grant agreement number 875171. AB is supported by a Cancer Research UK programme (Grant No. EPNCZS34).

7 | CONFLICT OF INTEREST

The authors declare no conflict of interest.

References

1. Li F, Zaslavsky AM, Landrum MB. Propensity score weighting with multilevel data. *Statistics in Medicine* 2013; 32(19): 3373–3387. doi: 10.1002/sim.5786
2. Schuler MS, Chu W, Co□man D. Propensity score weighting for a continuous exposure with multilevel data. *Health Services and Outcomes Research Methodology* 2016; 16(4): 271–292. doi: 10.1007/s10742-016-0157-5
3. Bates MD, Castellano KE, Rabe-Hesketh S, Skrondal A. Handling Correlations Between Covariates and Random Slopes in Multilevel Models. *Journal of Educational and Behavioral Statistics* 2014; 39(6): 524–549. doi: 10.3102/1076998614559420
4. Rabe-Hesketh S, Skrondal A. *Multilevel and longitudinal modeling using Stata*. Stata Press Publication . 2012.
5. Arpino B, Mealli F. The specification of the propensity score in multilevel observational studies. *Computational Statistics and Data Analysis* 2011; 55(4): 1770–1780. doi: 10.1016/j.csda.2010.11.008
6. Cole SR, Frangakis CE. The Consistency Statement in Causal Inference. *Epidemiology* 2009; 20(1): 3–5. doi: 10.1097/ede.0b013e31818ef366
7. Westreich D, Cole SR. Invited Commentary: Positivity in Practice. *American Journal of Epidemiology* 2010; 171(6): 674–677. doi: 10.1093/aje/kwp436
8. Greenland S, Robins JM. Identifiability, Exchangeability, and Epidemiological Confounding. *International Journal of Epidemiology* 1986; 15(3): 413–419. doi: 10.1093/ije/15.3.413
9. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology* 2009; 20(6): 880-883.
10. Chang TH, Stuart EA. Propensity score methods for observational studies with clustered data: A review. *Statistics in Medicine* 2022; 41(18): 3612–3626. doi: 10.1002/sim.9437

11. Austin PC, Lee DS. Estimating the Net Benefit of Improvements in Hospital Performance: G-Computation With Hierarchical Regression Models. *Medical Care* 2020; 58(7): 651–657. doi: 10.1097/mlr.0000000000001312

12. Skrondal A, Rabe-Hesketh S. Prediction in Multilevel Generalized Linear Models. *Journal of the Royal Statistical Society Series A: Statistics in Society* 2009; 172(3): 659–687. doi: 10.1111/j.1467-985x.2009.00587.x

13. Griswold ME, Zeger SL. On Marginalized Multilevel Models and their Computation. 2004; Johns Hopkins University, Dept. of Biostatistics Working Papers, Working Paper 99.

14. Tuerlinckx F, Rijmen F, Verbeke G, Boeck PD. Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology* 2006. doi: 10.1348/000711005x79857

15. Rizopoulos D. *GLMMadaptive: Generalized Linear Mixed Models using Adaptive Gaussian Quadrature*. 2023. https://drizopoulos.github.io/GLMMadaptive/, https://github.com/drizopoulos/GLMMadaptive.

16. Gibbons RD, Hedeker D, Charles SC, Frisch P. A Random-E□ects Probit Model for Predicting Medical Malpractice Claims. *Journal of the American Statistical Association* 1994; 89(427): 760–767. doi: 10.1080/01621459.1994.10476809

17. Farrell PJ, Macgibbon B, Tomberlin TJ. Bootstrap adjustments for empirical bayes interval estimates of small-area proportions. *Canadian Journal of Statistics* 1997; 25(1): 75–89. doi: 10.2307/3315358

18. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 2019; 38(11): 2074–2102. doi: 10.1002/sim.8086

19. Brumback BA, Dailey AB, Brumback LC, Livingston MD, He Z. Adjusting for confounding by cluster using generalized linear mixed models. *Statistics & Probability Letters* 2010; 80(21): 1650-1654. doi: https://doi.org/10.1016/j.spl.2010.07.006

Chapter 4

Causal inference for hierarchical (two-level) data: Inverse Probability-of-Treatment Weighting

4.1 Introduction

Since its introduction in 1983, the PS has been used to address both descriptive (in the case the 'treatment' is not manipulable, e.g. see application in [54]) and causal questions [73]; however, within the hierarchical framework, use of the PS is relatively new [5]. A few reasons might be: (i) compared to standard regression approaches, it may seem more complex to model the treatment first, and then use the PS either to weight, stratify, adjust or match on it - and this complexity may be magnified in presence of clustering that need to be addressed in the analyses; (ii) guidelines in the epidemiological literature are relatively scarce; (iii) balance has to be achieved (both on individual- and cluster-level confounders) when targeting the $p \quad ATE$); (iv) positivity issues that arise may seem more intricate to tackle within the multilevel setting [50]; (v) there is not a unifying, widely applied software to perform analyses which combine PS and two-level structures - in contrast to traditional regression adjustment methods.

In this chapter, we focus on how *IPTW* which models the treatment assignment mechanism, as well as its augmented version, *AIPTW* which models both the treatment and the outcome mechanisms, can be extended to the multilevel setting (see chapter 2 for the single level case).

Although there are several choices to account for clustering in the PS model (see 4.2), in this thesis, we focus on random effects models, because:

1. They are frequently applied in practice [66]; these models have been used to model the outcome and account for unmeasured cluster-level heterogeneity, however, in PS methods, the primary objective is balance on the individual and cluster characteristics, to obtain unbiased estimates; it is interesting to illustrate random effects performance when combined with PS weighting estimators.
2. Random effects models can be seen as an intermediate solution between fixed effects (which cease to be an alternative when there are too many different cluster levels and/or small cluster sizes) and marginal (on the clusters) models, which cannot be a choice when prior subject matter knowledge indicates important cluster-level confounders are omitted from the analysis.
3. There is interest in how to manage with the random effects distribution once having obtained the individual predicted probabilities of response (see chapter 3), which, for the PS model is the receipt of treatment. A frequent practice applied in recent studies is to directly plug in the Empirical Bayes Estimates (EBE) of the posterior modes of the random effects [48]; however, there is no explicit explanation for this choice, compared to other alternatives, such as integrating over the prior distribution of the random effects (see chapter 3). It would be interesting to examine the performance of the tested methods when either: (*i*) plugging in the EBEs or (*ii*) integrating over the prior distribution of the random effects when calculating the individual probabilities of treatment given the measured potential confounders.
4. Empirical evaluation of the relative performance of g-computation/standardization methods against PS weighting estimators may highlight different aspects of these estimators; at the same time, it would be interesting to focus on one modeling technique across analyses methods (i.e. random effects), to secure a fair comparison between the different methods.

Regarding focusing on weighting instead of stratifying, matching, or adjusting by the PS in two levels, for the last three, approaches have been examined in the multilevel setting [51, 99, 5]; nevertheless, weighting by the PS has desired properties (Chapter 2), while there is a gap in guidelines for IPTW practice within the two-level setting.

Although studies have demonstrated properties for IPTW estimators applied for multilevel data and continuous outcomes [53, 54], it is interesting to see whether similar

conclusions apply when targeting binary outcomes, which are very frequently encountered in cancer studies. For example, in Li et al., 2013, they write: *'...Analogous generalized linear models or generalized linear mixed models can be used for binary or ordinal outcomes...'* - however, it is not straightforward how these estimators behave in settings with non-continuous outcomes.

Finally, bias introduced when certain variables have been omitted from analysis has received little attention in the multilevel literature, with a focus mainly on other PS methods and other estimands, such as PS matching, PS covariate adjustment or stratification on the PS when targeting the ATT - see [51, 99, 5].

4.2 Inverse Probability-of-Treatment Weighting (IPTW) for two-level structures

Both IPTW and DR methods have in general two steps, including modeling the PS as a first step. Then, direct estimation of the ATE follows (as in IPTW) or a parametric model for the outcome fitted as a proper function of the baseline covariates X and the treatment Z (as in AIPTW). In a multilevel setting, the cluster component must be accounted for in the PS and/or the outcome modeling. Traditional estimators for the ATE as those in chapter 1 should account both for the individual- and cluster-level confounders to provide unbiased effect estimates. However we note that most often it happens to have important cluster-level covariates to be unmeasured in observational research. We now briefly present how IPTW (Chapter 1) can be extended to two-level designs [54]:

To give some intuition, let us refer to the expression of the $p \quad ATE$ [12], below:

$$E[\frac{ZY}{e(X)}] \quad E[\frac{(1 \quad Z)Y}{1 \quad e(X)}] \tag{4.1}$$

where, $e(X)$ is the true PS, expressed as a function of the pretreatment covariates X.

A natural extension to multilevel data would be to estimate $e(X)$ with a model that accounts for cluster-level confounders (either by including all confounders in the model or opting for models that account for clustering). The first case would be a marginal model (with respect to the clusters), and the second would be either a fixed or a random-effects model. We present these three possible modeling methods in what follows, although our focus is on random-effects modeling.

Let us now present the modeling options for a binary time-fixed treatment, assigned at the individual level as proposed by Li et al. in 2013 [54]. We also note that extensions for continuous treatments have also been suggested [75]:

i. Models for the PS

1. A marginal or single-level model:
   This model accounts for the clustered structure *only* through the *observed cluster-level covariates* (i.e, there is no independent variable in the model for specifying the cluster itself). As already mentioned, the PS is a function of the measured potential confounders $\mathbf{X}_{hk}$. For a binary treatment, a logistic regression may be fitted:

$$logit(e_{hk}) = d_0 + \mathbf{X}_{hk}a \tag{4.2}$$

   Where, $e_{hk} = Pr[Z_{hk} = 1|\mathbf{X}_{hk}]$, i.e., the individual PS and $\mathbf{X}_{hk} = (\mathbf{U}_{hk}; \mathbf{V}_h)$ the combined vector of both the individual- and cluster-level covariates, respectively. Lastly, d_0 is a fixed intercept.
   A model like (4.2) provides valid PS estimates if the unobserved covariates are independent of treatment assignment conditional on the observed covariates - a rather strong assumption in practice (as usually, there is unobserved residual confounding, either at the individual or at the cluster level - or, at both). The term *marginal* corresponds to the fact that it is not conditional on the cluster, making the estimation of the $p \quad ATE$ a straightforward procedure.
2. A fixed-effects model:
   This model includes a main fixed effect d_h that specifies the cluster. In this way, both the effects of observed and unobserved cluster-level covariates are captured. Thus, there is no need to include them as separate independent variables into the model. So, by including the fixed effect for the cluster level and the individual-level covariates we have:

$$logit(e_{hk}) = d_h + \mathbf{U}_{hk}a \tag{4.3}$$

   Where, $e_{hk} = Pr[Z_{hk} = 1|\mathbf{X}_{hk}; d_h]$, i.e., the individual PS and $\mathbf{X}_{hk} = (\mathbf{U}_{hk}; \mathbf{V}_h)$ the combined vector of both the individual- and cluster-level covariates, respectively. Lastly, d_h is a fixed intercept, indicative of cluster membership, with $h = 1; : : : H \quad 1$.

The fixed-effects model may result in larger variance estimates for the PS estimates than a correct model for the PS with fully observed cluster-level covariates $\mathbf{V}_h$. In the case of a great number of small clusters, it may provide unstable PS estimates, due to two reasons: i. large number of regression parameters to be estimated [1] - which, prohibits the use of fixed effects and ii. existence of clusters which contain only one of the two treatment groups. This positivity violation described in the latter case could be solved by excluding these clusters, at the cost of changing the estimand (ATE on a specific sub-population of clusters). Other solutions may be to group together clusters with similar characteristics, which, in the context of fixed effects shouldn't be applicable, as when clusters are represented by fixed effects, they are assumed as genuinely different entities; this however, is possible in the setting of random effects [50]. This leads to the next and focal modeling choice.

3. A random-effects model:
   This model is augmented with a prior distribution on the cluster-specific main effects. We focus on settings where this is the normal distribution, which is the most frequent choice in practice - however, there are other options as well [2]. If we assume a random intercept model for instance, where $d_{0h} \quad N(0; s_d^2)$, we may write:

$$logit(e_{hk}) = d_{0h} + \mathbf{X}_{hk}a \tag{4.4}$$

   Where, $e_{hk} = Pr[Z_{hk} = 1 j \mathbf{X}_{hk}; d_{0h}]$, i.e., the individual PS and $\mathbf{X}_{hk} = (\mathbf{U}_{hk}; \mathbf{V}_h)$ the combined vector of both any individual- and cluster-level covariates, respectively.
   This is preferred to a fixed-effects model, when there are many small clusters. Assuming that d_{0h} is a random variable which follows a distribution, diminishes substantially the number of regression parameters.
4. Other examples of models proposed in the literature, are surrogate indicator models [53], hierarchical modeling in the Bayesian framework [28] and generalized additive models [95].

To this point, we have presented potential modeling techniques for the PS model that account for clustering. If we recall expression (4.1), we may think that, to estimate the $p \quad ATE$, one may either construct a marginal (naïve) estimator,

[1] an example of the so-called Neyman-Scott incidental parameter problem.
[2] see for example, Cox regression with gamma shared frailty in 15.9 of [67].

which does not account for clustering, or, may, instead, construct an IPTW estimator, specific to each cluster, to first calculate the ATE within cluster (denoted as ATE_h for cluster h), and, finally, average over clusters to obtain an estimate for the $p \quad ATE$, which we target. Let us now present PS weighting estimators to estimate the $p \quad ATE$:

ii. Estimators for the $p \quad ATE$

1. A non-parametric marginal estimator [54]:

$$\hat{p}^{ma} = \sum_{Z_{hk}=1} \frac{Y_{hk}\hat{w}_{hk}}{\hat{w}_1} \quad \sum_{Z_{hk}=0} \frac{Y_{hk}\hat{w}_{hk}}{\hat{w}_0} \tag{4.5}$$

Where $\hat{w}_{hk} = 1=\hat{e}_{hk}$ for $Z_{hk} = 1$, $\hat{w}_{hk} = 1=(1 \quad \hat{e}_{hk})$ for $Z_{hk} = 0$ and $\hat{w}_z = \sum_{h;k:z_{hk}=z} \hat{w}_{hk}$ for $z = 0;1$. This is the traditional marginal IPTW estimator, i.e. the sum of the products of the individual-specific weights w_{hk} and the observed outcomes Y_{hk}, divided by the sum of the individual weights w_{hk}. This is calculated for the treated and, respectively the untreated. Normally, the difference provides the $p \quad ATE$. Therefore, this estimator includes no special step to account for the clustered structure. If both individual- and cluster-level confounders have been accounted for in the PS model, then balance on both should be achieved and unbiased $p \quad ATE$ estimates may be obtained. However, bias results from a simulation study [54] indicated observed relative bias [3] reached $\quad 20\%$ when the applied PS model was a random effects model - even within the analysis models that corresponded to the true model, referred to as the benchmark models. This result is in line with the fact that the random effects in the PS model cannot ensure balance on the cluster-level confounders, making them reliant on the inclusion of important cluster-level confounders.

2. A non-parametric clustered estimator [54]:
First, we estimate the ATE within each cluster:

$$\hat{p}_h = \frac{\sum_{k2h}^{z_{hk}=1} Y_{hk}\hat{w}_{hk}}{\hat{w}_{h1}} \quad \frac{\sum_{k2h}^{z_{hk}=0} Y_{hk}\hat{w}_{hk}}{\hat{w}_{h0}} \tag{4.6}$$

Where $\hat{w}_{hz} = \sum_{k2h}^{z_{hk}=z} \hat{w}_{hk}$ for $z = 0;1$. And finally:

[3] we have transformed to the equivalent of the relative bias here, by dividing the absolute bias by the true $p \quad ATE$ of the study; however, in the original study, the absolute bias results were reported (see Tables *I* and *II* in the original paper.

$$\hat{\rho}^{cl} = \frac{\sum_h \hat{w}_h \hat{\rho}_h}{\sum_h \hat{w}_h} \tag{4.7}$$

where $\hat{w}_h = \sum_{k2h} \hat{w}_{hk}$.

Thus, in the first step, we weight by the cluster-specific weights. In this way we obtain H different (as many as the clusters) cluster-specific estimators. The final estimator will be the weighted average of these, with weights the total weights within each cluster.

Lastly, we note that we refer to both of the aforementioned estimators as *non-parametric*, in the sense that they apply the observed outcomes to estimate the $p \quad ATE$.

4.3 Augmented Inverse Probability-of-Treatment Weighting (AIPTW) for two-level structures

Similarly to section 4.2, to give some intuition, let us refer to the expression of the $p \quad ATE$ [12], below:

$$E[\frac{ZY}{e(X)} \quad \frac{Z \quad e(x)}{e(x)} E_x[E(YjX;Z=1)]] \quad E[\frac{(1 \quad Z)Y}{1 \quad e(X)} \quad \frac{(1 \quad Z) \quad (1 \quad e(x))}{1 \quad e(x)} E_x[E(YjX;Z=0)]] \tag{4.8}$$

In the same manner as for IPTW, a natural extension in the two-level setting would be to simply estimate $e(X)$, $E(YjX;Z=1)$ and $E(YjX;Z=0)$ by a model that accounts for clustering. The modeling choices are identical to the ones presented for the PS in the previous section.

Therefore, we may obtain the parametric AIPTW estimator [54]:

$$\hat{\rho}^{dr} = \sum_{h;k} \hat{\rho}_{hk} n \tag{4.9}$$

with n being the total sample size, i.e., $n = \sum_{h=1}^{H} \sum_{k=1}^{n_h}$ and

$$\hat{\rho}_{hk} = \frac{Z_{hk} Y_{hk}}{\hat{e}_{hk}} \quad \frac{(Z_{hk} \quad \hat{e}_{hk})\hat{Y}^1_{hk}}{\hat{e}_{hk}} \quad \left[\frac{(1 \quad Z_{hk})Y_{hk}}{1 \quad \hat{e}_{hk}} + \frac{(Z_{hk} \quad \hat{e}_{hk})\hat{Y}^0_{hk}}{(1 \quad \hat{e}_{hk})} \right] \tag{4.10}$$

where, $\hat{e}_{hk}$, the estimated individual propensity score and $\hat{Y}^1_{hk}$, $\hat{Y}^0_{hk}$, the predicted individual response as if everyone were treated or untreated, respectively; the latter two can be obtained from any of the prediction options A, B or C mentioned in 3.2.1.

It can be proven that the estimator in (4.9) provides consistent $p \quad ATE$ estimates if either the treatment or the outcome model (or, both) is specified correctly. Although some authors have referred to this estimator as *parametric* [54], we refer to it as *semi-parametric*. This is in the sense that if the outcome model is incorrect, the observed (PS weighted) outcomes will be used (*non-parametric*). On the other hand, if the PS model is incorrect, then the fitted outcomes will be used to estimate the $p \quad ATE$ (*parametric*). Hence, we adopt the term *semi-parametric* instead [71].

The DR estimator proposed here, takes into account any unmeasured cluster-level characteristics that are independent of any of the measured unit- and cluster-level covariates by applying random effects to both the PS and the outcome model. In case of correlation of a cluster-level covariate with some included unit-level covariate, it is anticipated to introduce bias in the outcome model (conditional on the cluster) coefficient estimates. If that covariate is not the treatment, then, the same is anticipated for the PS model as well. It is known that within the single-level setting, when both the treatment and the outcome models are incorrectly specified, then the AIPTW estimator gives even more biased estimates compared to, e.g., an IPTW estimator with misspecified PS model. This is illustrated within our simulation scenarios for the two-level setting in 4.4.

Another point is that, version (4.9) does not calculate the cs-ATE first, to then average over clusters (like the IPTW non-parametric clustered version in (4.7)). So, in that manner, (4.7) could be seen as the clustered version of (4.5), while (4.9) could be seen as a doubly robust (in terms of outcome or PS model misspecification) version of (4.5).

4.3.1 Estimation of variance

Standard errors (SEs) for all the above estimators, can be derived via the Delta method. For the DR AIPTW estimator (4.9), we may write [54]:

$$(s^{dr})^2 = \mathring{a}_{h;k} (\hat{p}_{hk} \quad \hat{p}^{dr})^2 / n \tag{4.11}$$

with n being the total sample size, i.e., $n = \mathring{a}_{h=1}^{H} \mathring{a}_{k=1}^{n_h}$. However, this approach does not account for the uncertainty in the individual PS estimates.

Two alternatives, which instead take into account the estimation of the PS, are (*i:*) the non-parametric bootstrap, by sampling the clusters with replacement, and (*ii:*) the M-estimation technique - i.e., by stacking the required estimating equations to derive each of the estimators [57], [81].

4.4 Simulation study to evaluate the performance of standardization and PS weighted estimators for two-level structures

Herein, we present a numerical experiment we conducted to evaluate different versions of standardization and IPTW estimators performance under different scenarios of two-level data structures when targeting the *population or marginal* ATE ($p \quad ATE$). In particular [60]:

Aim

We evaluated and compared the performance of different standardization and IPTW estimators for two-level observational data in terms of bias under:

i. omission of a cluster-level covariate (across Figures 4.1, 4.2, 4.3, cluster-level covariate V_h is assumed non-observed),
and/or

ii. between-cluster variation in the effect of an individual-level covariate (on either the treatment or outcome),
and/or

iii. correlation between the cluster-level and an individual-level covariate is introduced (see Figures 4.2 and 4.3, for settings where V is assumed a confounder of the treatment and the outcome, or, predicts only the outcome, correspondingly).

The assumed cluster-level covariate V, may be either a predictor of the outcome or a confounder of the treatment and the outcome (see Figures 4.1, 4.2, 4.3).

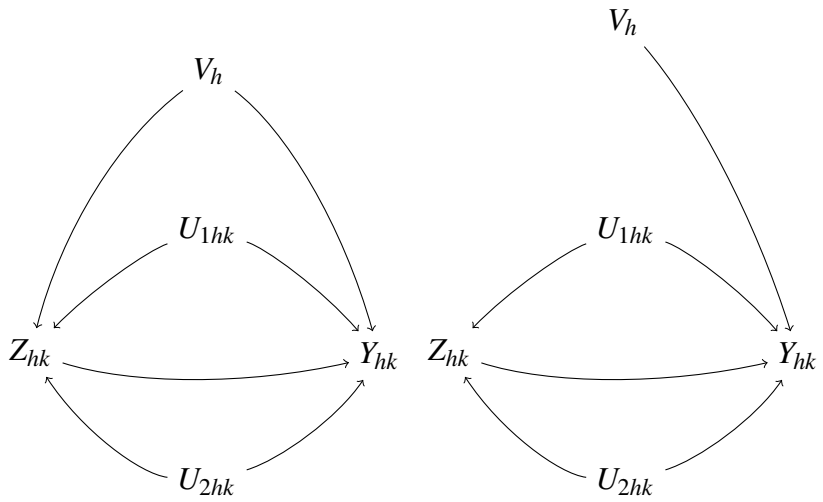


Fig. 4.1 DAG from left to right-hand side: V_h assumed *i:* confounder of Z_{hk} and Y_{hk}, *ii:* predictor of Y_{hk}; no correlation between V_h and U_{1hk}.

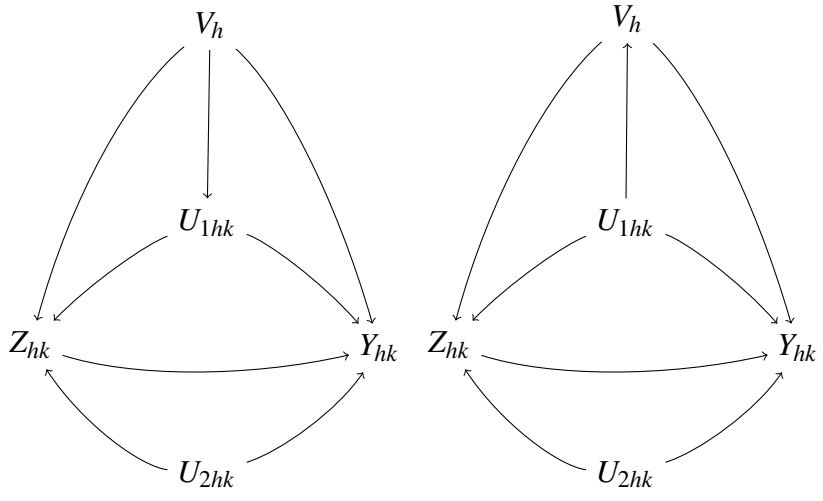


Fig. 4.2 DAG from left to right-hand side: V_h confounder of Z_{hk} and Y_{hk}; correlation between V_h and U_{1hk} could be translated with either V_h affecting or being affected by U_{1hk}.

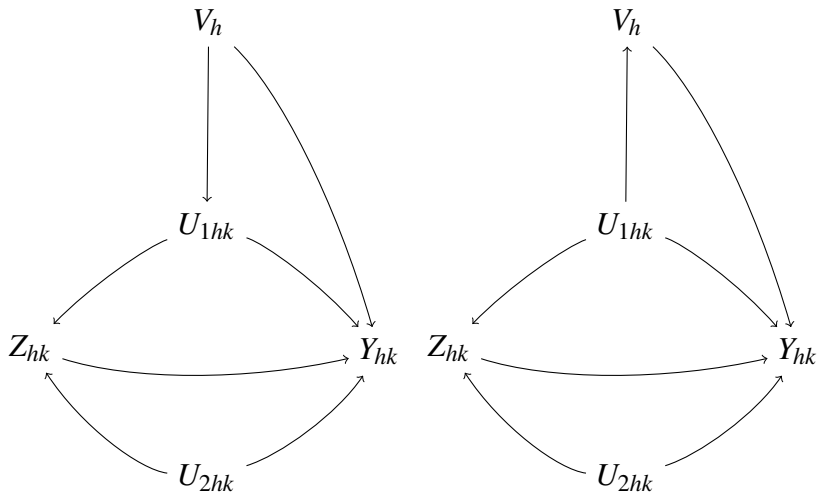


Fig. 4.3 DAG from left to right-hand side: V_h predictor of Y_{hk}; correlation between V_h and U_{1hk} could be translated with either V_h affecting or being affected by U_{1hk}.

Data-generating mechanisms and estimands

We randomly drawn a sample of $n_{total} = 1{;}000$ patients for a number of $n_{sim} = 1{;}000$ iterations. In particular, we simulated two individual-level confounders, $U_1 \quad N(0{;}1)$ and $U_2 \quad Bernoulli(0{:}5)$ and one cluster-level covariate, $V \quad N(0{;}1)$. These were assumed to be generated either (i) independently, or, (ii) with U_1 to be correlated with V, with a correlation, r and U_1 defined as: $U_1 := rV + K\sqrt{1 - r^2}$, with $K \quad N(0{;}1)$. Cluster-level covariate V, was allowed to be affecting either (i) only the outcome Y, or (ii) both the treatment Z and the outcome Y.

Values for the binary treatment Z, were generated from a $Bernoulli(e_{hk})$ for each patient (where, $e_{hk} = Pr(Z_{hk} = 1|\mathbf{X}_{hk}; a)$, the individual PS, and $\mathbf{X}_{hk}$ the vector that includes either only the individual-level $(U_{1hk}; U_{2hk})$ or both the individual- and cluster-level covariates, $(U_{1hk}; U_{2hk}; V_h)$), with:

(a) $e_k = expit(a_0 + a_1U_{1k} + a_2U_{2k})$, where a_0 a fixed intercept and $expit(x) = \frac{exp(x)}{1+exp(x)}$, or,

(b) $e_{hk} = expit((a_0 + d_{0h}) + a_1U_{1hk} + a_2U_{2hk} + a_3V_h)$, where $d_{0h} \quad N(0; s_d^2)$ is a random intercept, or,

(c) $e_{hk} = expit((a_0 + d_{0h}) + (a_1 + d_{1h})U_{1hk} + a_2U_{2hk} + a_3V_h)$, where d_{0h} and d_{1h}, a random intercept and random slope, respectively, which follow a multivariate (bivariate) normal distribution, i.e.:

$$\begin{pmatrix} d_{0h} \\ d_{1h} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \Sigma_d \right),$$

with Σ_d the variance-covariance matrix of the random intercepts and slopes, $Var(d_{0h}|\mathbf{X}_{hk}) = Var(d_{1h}|\mathbf{X}_{hk}) = 0.25$
and $r_{10} = r_{01} = 0.3$.

Then, a binary outcome was generated for each patient from a $Bernoulli(p_{hk})$ (where $p_{hk} = Pr(Y_{hk} = 1|Z_{hk} = z; \mathbf{X}_{hk} = \mathbf{x}; h)$, the individual probability of having the event conditional on the treatment, covariate and random effects values - if the latter is assumed), with:

(a) $p_{hk} = expit(\beta_0 + \beta_1 U_{1hk} + \beta_2 U_{2hk} + \beta_3 V_h + \beta_4 Z_{hk})$, where β_0 a fixed intercept, or,

(b) $p_{hk} = expit((\beta_0 + h_{0h}) + \beta_1 U_{1hk} + \beta_2 U_{2hk} + \beta_3 V_h + \beta_4 Z_{hk})$, where $h_{0h} \sim N(0; 1)$ is a random intercept, or,

(c) $p_{hk} = expit((\beta_0 + h_{0h}) + (\beta_1 + h_{1h})U_{1hk} + \beta_2 U_{2hk} + \beta_3 V_h + \beta_4 Z_{hk})$ where h_{0h} and h_{1h}, a random intercept and random slope, respectively, that follow a multivariate (bivariate) normal distribution:

$$\begin{pmatrix} h_{0h} \\ h_{1h} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \Sigma_h \right) \tag{4.12}$$

where Σ_h the variance-covariance matrix of the random intercepts and slopes, $Var(h_{0h}|\mathbf{X}_{hk}) = Var(h_{1h}|\mathbf{X}_{hk}) = 0.25$,
and $r_{10} = r_{01} = 0.3$.

We allowed the following parameters to vary across simulation scenarios:

1. Treatment prevalence: $p_Z = Pr(Z = 1) \in \{0.20; 0.50\}$
2. Marginal event rate in the controls: $p_{Y_0} = Pr(Y_0 = 1) \in \{0.10; 0.30\}$
3. Correlation, r, between the cluster-level covariate, V, and the individual-level covariate, U_1: $r \in \{0; 0.5\}$
4. Number of clusters, H and average cluster size, n_h for an unbalanced design (we kept the total sample size constant and equal to 1,000 patients); we applied an

unbalanced design, which is more complex than the balanced equivalent, and, more plausible to be encountered in real data sets:

$$(H; n_h) \in \{(50; 20); (100; 10)\}$$

5. true *marginal* or *population* ATE ($p \quad ATE$), expressed as the risk difference (RD). As the risk difference is an absolute measure of association, its value cannot reflect the true intensity of the treatment effect, as a small risk difference could equally likely represent a mild, medium or strong true effect, depending on the values of the risk under treatment and control levels, respectively. For that reason, we translated our assumed risk differences to the corresponding marginal risk ratio, being $RR \in \{1; 0.8\}$ - i.e., null or moderate (for the latter, we approximated the true marginal treatment effect by calculating it independently for each DGM, from a sufficiently large number of observations, often referred to as a *superpopulation* - e.g., $n = 10^7$; more specifically, we set the marginal risk ratio (RR) to be equal to 0.8, expressing a moderate effect - as the RR is a relative measure as opposed to the RD). The same value assigned for different event rates may or may not be moderate, e.g.: if we have a true event rate of 5%, then a 3% ATE is not just moderate, whereas, for an assumed event rate of 50%, it is negligible.

   Then, for each different scenario of event rate under control, we calculated the required event rate under treatment - e.g., for a true event rate of 10% for the control group, we calculated the respective p_{Y_1} that would provide a risk ratio of 0.8, which is equal to 0.08. In that case, the corresponding risk difference would be -0.02; via trial and error we set the required coefficient values in our data generating models to obtain sample risk differences that were close to -0.02).

   Overlap of the PS distributions between the two treatment groups was graphically checked across simulated data sets and DGMs, as this is affected by the assumed treatment prevalence of each simulation scenario. Alternatively, other authors set the assumed overlap in advance, as in [100]. Correlation between random intercepts and random slopes was set to be moderate and equal to 0.3.

   As for the number of clusters where only one treatment level is assigned, we did not set this in advance, as it would be a natural consequence of the design (e.g., for low treatment prevalence, it is intuitive to anticipate a large number of clusters, within which, only one treatment level is assigned - however, in our scenarios the chance of receiving treatment was either moderate or equal to not receiving it);

the proportions of clusters with only one treatment level being assigned may be calculated for each simulated dataset within each DGM [4].

For each paired combination of PS and outcome model, a full factorial design corresponds to a total of 2 2 2 2 2 3 3 = 288 data-generating mechanisms. However, hereafter (see 4.1 and Appendix C) we present results related to a total of 207 DGMs, with the remaining not presented in this dissertation due to time restrictions[5]. To perform the simulations and analysis methods, we used the multiple purpose statistical software **R**, version *4.3.0*; we applied the packages ICCbin, boot, MASS, statmod, foreach, GLMMadaptive, lme4, rsimsum, dplyr, ggplot2 and xtable. Due to the computational power required both to generate and analyze the simulated data sets, we were granted access to LSHTM's High Performance Computing (HPC) cluster [6]. Reproducible R code to perform the simulations and analyze the data can be found on GitHub at sims-ISCB2023.

Setting an unbalanced design

Across DGMs we applied an unbalanced design of 50 clusters of an average of 20 patients within each cluster. We assumed the cluster sizes to be following a multinomial distribution. To set the desired range of cluster sizes across clusters, we applied the same procedure as in Appendix A of [19]. Relevant code to generate the cluster sizes can be found on GitHub, at sims-ISCB2023.

Setting the required parameter values

We now briefly describe the empirical procedure we applied to obtain the required sets of parameters that provided to a satisfactory degree the assumed marginal event rates for the control group p_{Y_0}, treatment prevalence p_Z and moderate effect of the marginal ATE ($p \quad ATE$) (i.e., a marginal RR equal to 0.8). After having obtained the desired

[4] all the simulated data sets were stored, with information of the treatment level received for each individual observation within each cluster, corresponding to each simulated dataset across DGMs.

[5] 81 DGMs corresponding to codes 3,4,12,14,15,16,19,20,31,32,36,37,38,39,44,47,48,50,51,52,68,69,70,79,80, 85,94,96,109,113,119,121,124,125,134,139,141,148,158,164,182,186,193,197,201,202,205,206,209,214,217, 218,221,222,224,226,234,237,239,241,245,249,250,253,254,259,261,262,264,266,269,270,271,274,277,278, 280,281,282,285,286 were not included in *this version of the simulation results* presented in 4.4 and in Appendix C.

[6] Rocky Linux 8: 14 nodes with between 320GB-1024GB RAM and 24-64 CPU cores (a mixture of Intel and AMD CPUs), and 1 GPU node with 2 Nvidia A40 GPUs. Each user has 200GB of storage by default, more can be requested Slurm provides both single node multi core/processor parallel computation and MPICH for parallel computation across the cluster.

sets of parameter values, we applied graphical checks of overlap of the PS distributions between the two treatment groups across clusters under the parameter values that we set [7].

We began by setting fixed intercept values, a_0, b_0, for the PS and outcome models respectively, in order to obtain the corresponding p_Z and p_{Y_0}. Set values for p_{Y_0} directly lead to certain p_{Y_1}, to acquire the assumed moderate marginal ATE (as we assumed $\frac{p_{Y_1}}{p_{Y_0}} = 0.8$). To shift the p_{Y_1} to the desired one, we applied different values for the treatment coefficient of the outcome model, b_4:

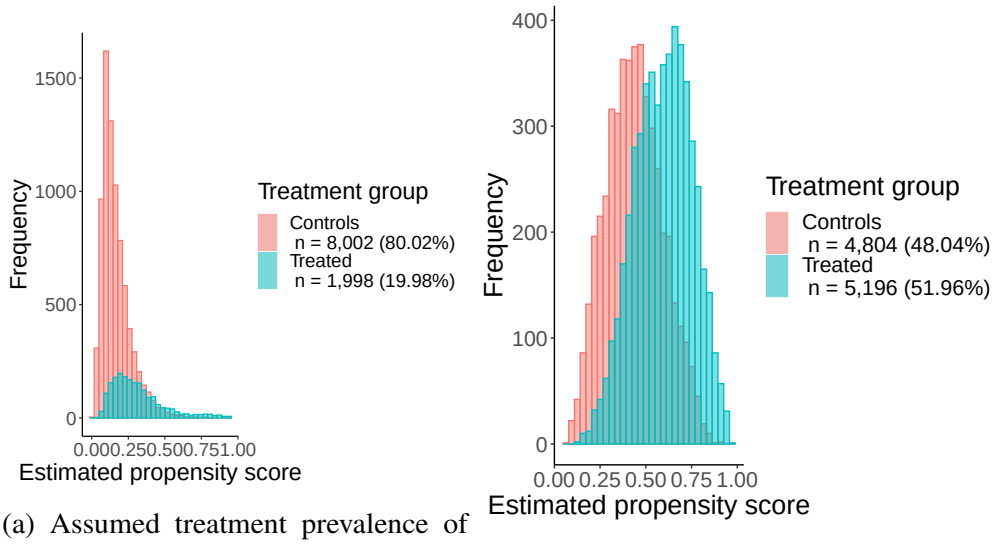
- For $p_Z = 0.20$, we set $a_0 = \quad 1.499$;
- For $p_Z = 0.50$, we set $a_0 = 0.15$;
- For $(p_{Y_0}, p_{Y_1}) = (0.10, 0.08)$, we set $(b_0, b_4) = (\quad 2.35, \quad 0.31471)$;
- For $(p_{Y_0}, p_{Y_1}) = (0.30, 0.24)$, we set $(b_0, b_4) = (\quad 0.63, \quad 0.41)$.

We generated a sample of $10,000$ observations in total (500 clusters of 20 patients on average within each cluster). We iterated $1,000$ times and checked whether the estimates for the different data-generating models were close enough to the assumed values for p_Z, p_{Y_0}, p_{Y_1} (e.g., arbitrarily, we assumed that precision up to the third decimal point should suffice). We added a layer of checks by drawing a superpopulation of $10,000,000$ observations in total (i.e. 5 $\quad$ $100,000$ clusters of 20 patients on average within each cluster), for which case it may be assumed parameter estimates $\hat{p}_Z, \hat{p}_{Y_0}, \hat{p}_{Y_1}$ approximate well the true desired parameter values for p_Z, p_{Y_0}, p_{Y_1}.

To obtain the desired level of overlap (we assumed sufficient overlap in our settings), we started from an initial set of parameter values for the coefficients of the PS model that corresponded to the effects of U_1, U_2 and V, namely (a_1, a_2, a_3).

We then checked the level of overlap graphically, by plotting the PS distributions of the two treatment groups (Figures 4.4, 4.5) in a sample of 10^7 observations in total. By trial and error we fixed (a_1, a_2, a_3) to be equal to $(0.1, \quad 0.2, 0.6)$. All the aforementioned values worked across the different assumed PS and outcome data generating models (i.e., marginal, random intercept and random slope models), as, even in the random effects case we assumed mean values of zero for the random effects, and no between effect interactions throughout the models.

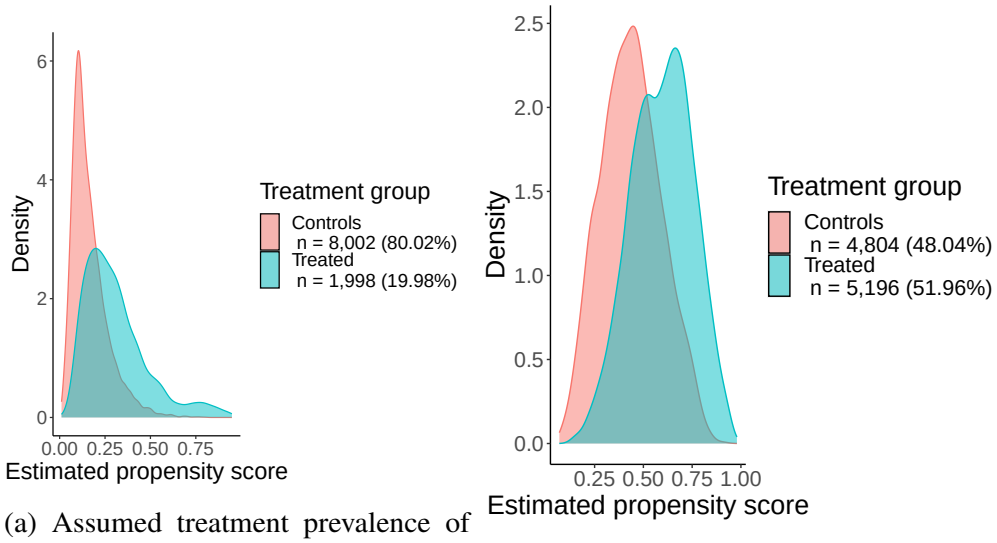
[7] however, since we are applying estimators which are calculating the ATE within cluster first (i.e., non-parametric clustered estimators), it would be ideal to check overlap within clusters too - we assumed this was sufficient by design, as the number of clusters with only one treatment level being assigned was assumed very low).



(a) Assumed treatment prevalence of 20%.

(b) Assumed treatment prevalence of 50%.

Fig. 4.4 Histograms of the estimated propensity score distributions among those under treatment versus among those under control; random draws of 10,000 observations.



(a) Assumed treatment prevalence of 20%.

(b) Assumed treatment prevalence of 50%.

Fig. 4.5 Density plots of the estimated propensity score distributions among those under treatment versus among those under control; random draws of 10,000 observations.

Methods of analysis

Methods to analyse the simulated data sets are presented in Table 4.1. We note that analysis of both outcome models with no random effects and outcome models with random intercepts was performed by random intercept models. Additionally, analysis of random slope outcome models was performed by random slope models across the different tested simulation scenarios. That could potentially favour results across random effects DGMs *vs.* DGMs with no random effects for the outcome.

Table 4.1 Summary of analysis methods[†].

Analysis method	**Description**
Bench EBE[‡]	True outcome model - individual estimated probabilities of outcome conditional on Empirical Bayes Estimates of random effects
Bench PA[□]	True outcome model - individual estimated probabilities of outcome averaged over the random effects distribution
G-comp EBE	G-computation Empirical Bayes estimates
G-comp PA	G-computation Population average estimates
IPTW-Mar EBE	IPTW-Marginal estimator Empirical Bayes estimates
IPTW-Mar PA	IPTW-Marginal estimator Population average estimates
IPTW-Cl EBE	IPTW-Clustered estimator Empirical Bayes estimates
IPTW-Cl PA	IPTW-Clustered estimator Population average estimates
AIPTW EBE	AIPTW Empirical Bayes estimates
AIPTW PA	AIPTW Population average estimates

[†] V omitted across analysis methods - excluding benchmark models.

[‡] Empirical Bayes Estimates.

[□] Population Average.

Performance measures

Herein, we present results for bias estimates, i.e., $\frac{1}{n_{sim}} \mathring{a}_{i=1}^{n_{sim}} \hat{q}_i \quad q$, as this was our primary objective [8]. Other performance measures that are usually examined are: coverage, root mean squared error, empirical standard error, model based standard error, convergence, type I error rate, and their respective Monte Carlo standard errors, to account for uncertainty across simulated samples [60].

[8] note that in our case $q = ATE$.

Results

One may fit a regression of the performance measure of interest (e.g., bias) on the varying parameters of the simulation design, to get an idea of which parameters are significant for comparison of the analysis methods (see Tables 4.2, 4.3, 4.4, 4.5). Briefly, both across methods and within the IPTW analysis methods (Tables 4.2 and 4.4): b_0, i.e., the true event rate in the controls, r, i.e., the true correlation between U_1 and V, RR, i.e., the true risk ratio, and additionally the true treatment model and the number of clusters (or, equivalently, the average cluster size[9]), were significant at 5% level of statistical significance. Within AIPTW analyses (Table 4.5), **additionally**, a_0 (which determines the true treatment prevalence) was significant at 5% level of statistical significance, with the true outcome model remaining non-significant. Within the benchmark and g-computation analysis methods (Table 4.3), a_0 and r were no longer significant, whereas the true outcome model became significant.

Table 4.2 Coefficient estimates, standard errors and corresponding p-values for linear regression of estimated bias fitted on the simulation design parameters across analysis methods; statistically significant at 5% significance level.

term	estimate	std.error	p.value
intercept	-0.030	0.006	0.000
method	0.004	0.000	0.000
a_0	-0.001	0.001	0.326
b_0	0.008	0.001	0.000
r	0.010	0.002	0.000
RR	0.019	0.006	0.001
treatment model random intercept	0.022	0.001	0.000
treatment model random slope	0.022	0.001	0.000
outcome model random intercept	-0.001	0.001	0.428
outcome model random slope	-0.001	0.002	0.441
number of clusters	0.000	0.000	0.028

[9] as the total number of observations within our settings was constant.

Table 4.3 Coefficient estimates, standard errors and corresponding p-values for linear regression of estimated bias fitted on the simulation design parameters within g-computation and benchmark analysis methods; statistically significant at 5% significance level.

term	estimate	std.error	p.value
intercept	-0.029	0.002	0.000
method	0.006	0.000	0.000
a_0	0.000	0.000	0.293
b_0	0.001	0.000	0.001
r	0.000	0.001	0.769
RR	0.013	0.002	0.000
treatment model random intercept	0.011	0.001	0.000
treatment model random slope	0.010	0.001	0.000
outcome model random intercept	-0.001	0.001	0.024
outcome model random slope	-0.001	0.001	0.117
number of clusters	0.000	0.000	0.000

Table 4.4 Coefficient estimates, standard errors and corresponding p-values for linear regression of estimated bias fitted on the simulation design parameters within IPTW analysis methods; statistically significant at 5% significance level.

term	estimate	std.error	p.value
intercept	0.095	0.009	0.000
method	-0.015	0.001	0.000
a_0	-0.001	0.001	0.377
b_0	0.009	0.001	0.000
r	0.009	0.003	0.006
RR	0.020	0.008	0.013
treatment model random intercept	0.024	0.002	0.000
treatment model random slope	0.023	0.002	0.000
outcome model random intercept	-0.001	0.002	0.754
outcome model random slope	-0.001	0.002	0.764
number of clusters	0.000	0.000	0.426

Let us now present results for the bias estimates that correspond to DGMs of a random intercept and slope treatment model and a random intercept outcome model, under the assumption of (i) no correlation and (ii) correlation between U_1 and V across different scenarios of event rates for the controls, p ATE value and combinations of unbalanced cluster designs - see Figures 4.6 to 4.7). These results highlight a prevailing pattern across the assumed 288 DGMs and methods examined. Hereafter, we summarize some of the main findings:

Table 4.5 Coefficient estimates, standard errors and corresponding p-values for linear regression of estimated bias fitted on the simulation design parameters within AIPTW analysis methods; statistically significant at 5% significance level.

term	estimate	std.error	p.value
intercept	-0.333	0.018	0.000
method	0.036	0.002	0.000
a_0	-0.002	0.001	0.034
b_0	0.019	0.001	0.000
r	0.033	0.003	0.000
RR	0.027	0.009	0.002
treatment model random intercept	0.043	0.002	0.000
treatment model random slope	0.041	0.002	0.000
outcome model random intercept	-0.001	0.002	0.489
outcome model random slope	-0.002	0.002	0.299
number of clusters	0.000	0.000	0.035

- Both of the benchmark models estimates were unbiased across DGMs as they described the true DGMs[10] (Figures 4.6 and 4.7).
- Standardization or g-computation estimates of the $p \quad ATE$ seemed less biased compared to their counterpart marginal IPTW performance - irrespective of whether we averaged over the random effects distribution or plugged-in the EBEs to predict the individual probabilities of response. That was consistent across PS and outcome data generating models. In terms of modeling assumptions, when we model the outcome mechanism and omit the cluster-level confounder from the random effects model, we expect the independence of the random effects and the included covariates to be violated (either for scenarios of presence or absence of correlation between U_1 and V). On the other hand, when we model the treatment assignment mechanism, in terms of obtaining balance, the PS analysis model does not include the cluster-level confounder, V, hence, balance for the distribution of V between the two treatment groups cannot be achieved. To get unbiased estimates of the $p \quad ATE$ when applying IPTW, one needs to reassure balance for both the individual- and cluster-level characteristics (see Appendix E). This is, contrary to the clustered IPTW, for which balance is achieved for both the measured individual- and *any* cluster-level covariates - as this estimator, in a first step estimates the ATE for each cluster and it averages over clusters at a second step to estimate the $p \quad ATE$ from the sample. In our settings (i.e., under either null or moderate $p \quad ATE$), it seemed more important to achieve balance on both individual and cluster characteristics

[10] analyses of both the outcome models with no random effects and outcome models with random intercepts were performed by random intercept models; analyses of random slope outcome models were performed by random slope models across the different tested simulation scenarios.

to obtain unbiased $p \quad ATE$ estimates for the IPTW methods involved, than to reassure the independence between U_1 and V for the methods combining random effects when applying standardization/g-computation.

For scenarios where V was not a confounder, performance across methods was almost unbiased - although the marginal IPTW seemed to have slightly worse performance compared to the rest of the tested estimators - however, bias was still below 5% across DGMs (see complementary results in Appendix C).

– The clustered estimator had no bias, as anticipated, regardless of the analysis model used for the PS or the underlying DGM. This estimator, first estimates the $cs \quad ATEs$, and on a second step, averages over the $cs \quad ATEs$ to obtain the $p \quad ATE$. For this estimator, balance is ensured on any cluster-level confounder, hence, as a diagnostic test, only balance on the individual characteristics must be checked. We highlight that it is required all the causal assumptions are met within each cluster, including the positivity assumption; therefore, any cluster that consists of patients being assigned only one of the two treatment levels must be excluded; in the case of existence of many clusters like that, one has to consider the potential of re-defining their target population to the one that consists of the clusters that include both treatment levels (although this setting, is more likely to happen when we describe the different clusters by fixed effects for the cluster membership - i.e., not within our examined settings). In the literature, relatively recently, some suggested criteria of grouping similar clusters together and estimating the ATE for each group of clusters first, then average over the groups of clusters (which, can be seen as new clusters, having both treatment levels being assigned within them [50]). This solves the issue of having to exclude observations from the analysis.

– Interestingly, when looking at DGMs with a moderate $p \quad ATE$ and event rate of 30% in the controls (Figure 4.7), the bias estimates increased. In particular, for the marginal IPTW and the AIPTW methods, observed bias was nearly 10% in some cases. This may be attributed to the fact that the marginal IPTW does not account for clustering - unless the model for the PS does so. However, one would argue that in our settings, the PS model should be accounting for clustering, since it includes random effects. However, as already mentioned, the objective of the PS model, is to obtain balance across individual and cluster-level covariates after weighting. By including random effects in the PS model, balance cannot be necessarily achieved, as the random effects estimates tend to shrink towards the cluster mean [54]. For an event rate of 10% in the controls, bias patterns were similar to the respective plots for an assumed null ATE (not presented herein).

– When applying AIPTW, both the treatment and the outcome model were misspecified, therefore we assume that the low performance can be attributed to that (Figures 4.6 and 4.7) - as, in the single level setting we know that an AIPTW does not necessarily remain unbiased if a confounder is omitted from both the treatment and the outcome model [46].

- When looking at the population averaged versions of the evaluated analysis methods compared to their EBE counterparts: within g-computation/standardization, the performance was equally unbiased (Figures 4.6 and 4.7). Within IPTW-mar and AIPTW, the EBE versions performed better than their population averaged counterparts across DGMs (see also nested loop plots in Appendix C). One possible explanation on that could be the fact that when predicting the individual PS by averaging over the prior distribution of the random effects, balance on the cluster-level characteristics is not achieved. On the other hand, by plugging-in the EBEs in the individual predicted PSs, some "better" balance on the cluster-level confounder V, may be achieved - although, not optimal, as the EBEs tend to shrink towards the cluster mean.
- Finally, the above conclusions are consistent across DGMs that correspond to outcome models with no random effects (see Figures C.1, C.4 and C.7 in Appendix C) - for which, analysis models do not coincide with the assumed data generating models (as in the opposite case, the analysis models should be favoured in terms of estimating performance).

  So far, we have investigated how the different causal inference methods of choice perform, in terms of bias, in the two-level setting. In the next chapter, we present two different estimating methods that are used for large-sample variance estimation, and in particular, for settings where the variance formula is intricate to obtain, such as in IPTW estimation. Variance estimation of the following chapter, is examined in the single-level setting.

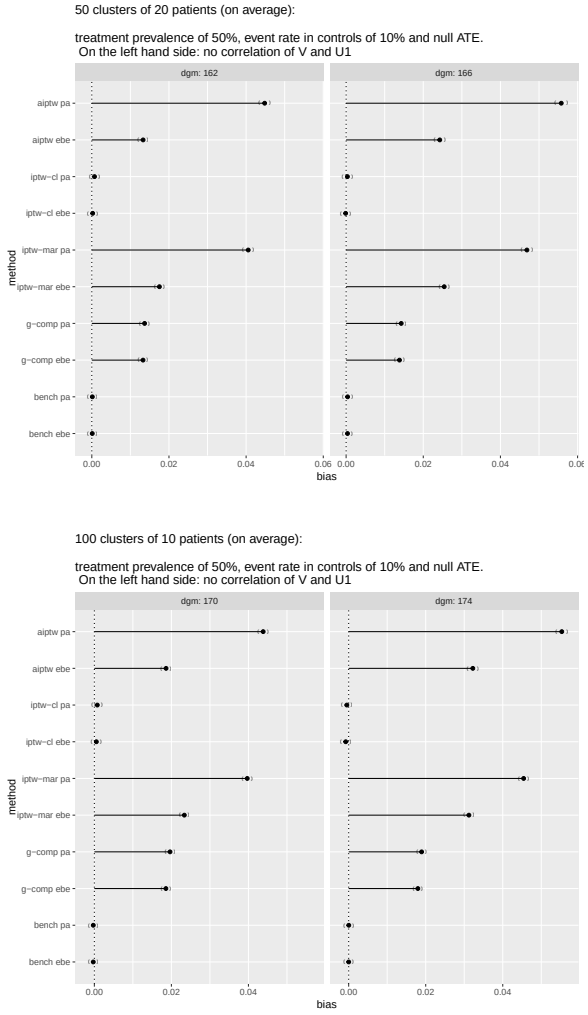


Fig. 4.6 Lollipop plot of bias (Monte Carlo 95% confidence intervals in parentheses): random intercept and slope treatment model and random intercept outcome model; 1% of models with convergence and/or singularity warnings were excluded from results.

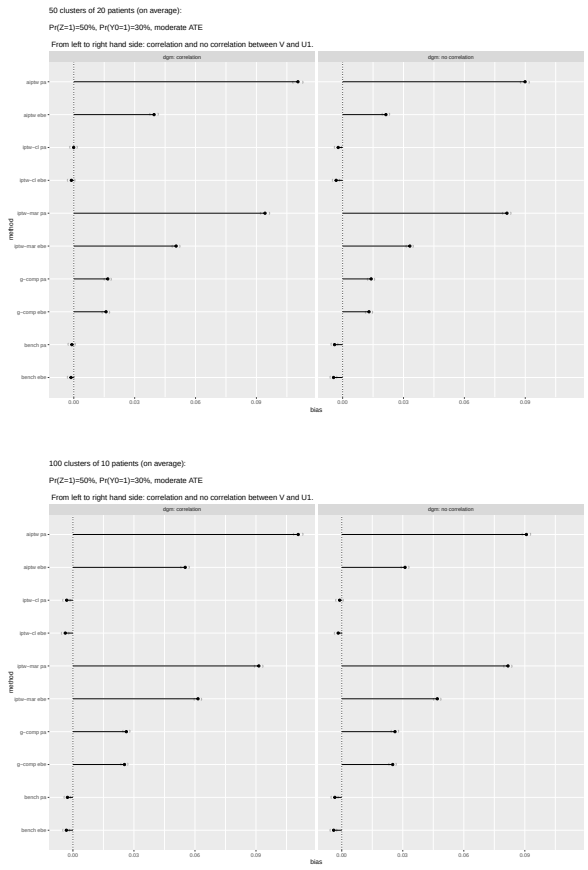


Fig. 4.7 Lollipop plot of bias (Monte Carlo 95% confidence intervals in parentheses): random intercept and slope treatment model and random intercept outcome model; 1% of models with convergence and/or singularity warnings were excluded from results.

Chapter 5

Unifying variance estimation for different estimands using IPTW in the single-level setting

5.1 Introduction

Herein, we present three general approaches to estimate the variance of a parameter estimator - which, is a parameter itself. These approaches are useful to apply when it is not straightforward to obtain a closed-form variance estimator. In order of appearance, these are: *M-estimation (or estimating equations)*, *linearization* and, *the bootstrap*. The first two provide approximate formulas, which work well in large-samples. The last one does not provide closed-form estimators for the variance, however it has the flexibility to work in various settings, likewise, when the sample size is *large enough*.
In our framework, we target to estimate the causal treatment effect q. For each estimator, $\hat{q}$, we also need to estimate its respective variance $Var(\hat{q})$ by $\widehat{Var}(\hat{q})$. As we will see, all three methods may be applied to estimate the variance of Inverse Probability-of-Treatment Weighted (IPTW) estimators unbiasedly, even when the target is an effect in various different populations, apart from the traditional ATE. These additional populations are the population of the treated (average treatment effect in the treated - ATT) and the population of those who have equal chance of receiving any of the treatment levels (average treatment effect in the overlap - ATO). Moreover, it can be proven both via M-estimation and linearization that we may obtain closed-form approximate variance formulas, which are general to the target population in question and also take into account the uncertainty in the PS estimates (see 5.2 and Appendix B). In addition, these formulas may be general to the choice of the PS model - simply by changing the score equation applied to obtain the respective regression parameter estimates for that model. In Section 5.2, we present a paper published in Statistics in Medicine that offers a comprehensive way to obtain these variance estimators analytically, along with corresponding R code for the application of the formulas. A brief description of a simulation study to illustrate

the performance of the different variance estimators across different target populations as well as elaborate analytical derivations of the formulas via M-estimation and linearization are provided in Appendix B.

5.1.1 M-estimation theory

This subsection is heavily based on [81] and Chapter 7 by Boos and Stefanski in [15].
M-estimation (or estimating equations), is a method to estimate any target parameter θ, by an estimator $\hat{\theta}$ which is called an *M-estimator*, if it satisfies the below:

$$\sum_{i=1}^{n} \psi(\mathbf{O}_i; \hat{\theta}) = 0 \tag{5.1}$$

where $\mathbf{O}_i$ are independent observations, θ is a k-vector of parameters to estimate and $\psi(\cdot)$ is a k-dimensional known function that does not depend on either i or n. In (5.1), $\mathbf{O}_i$ represents the i^{th} observation.
Sometimes, the above notation is adapted to accentuate the dependence of $\mathbf{O}_i$ on specific components. For example, when we regress the treatment Z to build a PS model, we observe $\mathbf{O}_i = (x_i; Z_i)$. Therefore, (5.1) becomes:

$$\sum_{i=1}^{n} \psi(x_i; Z_i; \hat{\theta}) = 0 \tag{5.2}$$

In our framework, IPTW estimators for the population means can be seen as what in the literature is called *partial M-estimators*. A partial M-estimator is an estimator that can be seen as an M-estimator as soon as certain unknown parameters are estimated. For instance, when targeting the ATE, for the population mean under treatment, the IPTW estimator is defined:

$$\hat{\mu}_1 = \left(\sum_{i=1}^{n} \frac{Y_i Z_i}{\hat{e}_i} \right) \Bigg/ \left(\sum_{i=1}^{n} \frac{Z_i}{\hat{e}_i} \right) \tag{5.3}$$

Therefore, to write $\hat{\mu}_1$ as an M-estimator, we need Y_i, Z_i, which are observed and the individual PS estimates $\hat{e}_i$. The last can be estimated by a model for the PS:

$$\sum_{i=1}^{n} \psi(x_i; Z_i; Y_i; \hat{\theta}) = 0 \tag{5.4}$$

If we choose to fit a generalised linear model (GLM) on the treatment, the previous becomes:

$$\begin{pmatrix} \sum_{i=1}^{n} (Y_i - \mu_1) Z_i \frac{1}{e_i} \\ \sum_{i=1}^{n} \frac{\mathbf{x}_i (Z_i - e_i)}{e_i (1 - e_i) g'(e_i)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{5.5}$$

while in eq. (5.4), $\hat{\theta} = (\hat{m}_1; \hat{\alpha}^\top)^\top$. The second "row" in (5.5) is the score equations to derive the regression parameter estimates $\hat{\alpha}$ obtained from the respective PS model. Note that $\mathbf{x}_i = (1; x_{i1}; \ldots; x_{ip})^\top$. Lastly, $g(\cdot)$ is the link function applied to the PS model (for example, a $logit(\cdot)$ for a binary treatment). More details on how to derive M-estimators for the population mean under no treatment and, finally the ATE, ATT and ATO, can be found in Appendix B.

How is the variance-covariance matrix defined It can be proven that $\hat{\theta}$ is asymptotically normally distributed, i.e., $\hat{\theta} \quad N(\theta_0; \Sigma_{\theta_0})$, where, Σ_{θ_0} is the variance-covariance matrix equal to $n^{-1}A^{-1}(\theta_0)B(\theta_0)A^{-T}(\theta_0)$, with $A(\theta_0) = -\mathsf{E}[\frac{\partial \psi}{\partial \theta^T}]$ and $B(\theta_0) = \mathsf{E}[\psi\psi^T]$.
In the previous example, where $\hat{\theta} = (\hat{m}_1; \hat{\alpha}^\top)^\top$, A and B are:

$$A = -\mathsf{E}\begin{bmatrix} Z\frac{1}{e} & 0 \\ 0 & \mathbf{x}\mathbf{x}^T e(1-e) \end{bmatrix}$$

and

$$B = \mathsf{E}\begin{bmatrix} (Y-m_1)^2 Z(\frac{1}{e})^2 & 0 \\ 0 & \mathbf{x}\mathbf{x}^T (Z-e)^2 \end{bmatrix}$$

Desired properties of M-estimators As mentioned above, it can be shown that M-estimators are consistent and asymptotically normal. For a proof, see Section 7.8 in [15]. Hence, by demonstrating that an estimator is an M-estimator, we simultaneously prove that the estimator in question is both consistent and follows a normal distribution, making it an appealing method to apply for parameter estimation. Methods widely applied, like the GEEs and the delta method can also be seen as M-estimators (see Sections 7.5.6 and 7.2.4 in [15], respectively).

5.1.2 Linearization technique

In linearization, we seek for an artificial variable $l_i(\hat{\theta})$, which we shall call the linearized variable of $\hat{\theta}$, to express the difference between the estimator $\hat{\theta}$ and the parameter θ, as a sum of $l_i(\hat{\theta})$, plus a negligible quantity, $O_p(n^{-1=2})$ — which approaches zero as the sample size increases at a rate of $1=\sqrt{n}$, where n is the sample size:

$$\hat{\theta} - \theta = \frac{1}{n}\sum_{i=1}^{n} l_i(\hat{\theta}) + O_p(n^{-1=2}) \tag{5.6}$$

Eq. (5.6) ensures that $Var(\hat{\theta}) \quad n^{-1}Var(l_i(\hat{\theta}))$. For more details refer to Section 4 of Paper in 5.2.
We note that historically, there are two fields in which linearization (and the influence curve) methods have been widely applied: (i) survey methodology [24] and (ii) robust statistics [37]. In our paper, we follow the definitions introduced in 1999 by Deville [24] within the survey methodology setting.

Linearization (influence curve) and connections to M-estimation It can be proven that there are connections between the linearized variable (or influence curve), $l_i(\hat{q})$ and the respective M-estimator, $\hat{q}$ (briefly, in Section 7.3 in [15], it is demonstrated that: $l(q_0) = \mathbf{A}(q_0)^{-1} y(O; q_0)$ for M-estimators, and therefore: $Var(l(q)) = Var(q_0)$). The linearization (or, influence curve) method is more general than the M-estimation approach; however, in many (as in our) settings, these are equivalent. Since stacking estimating equations and computing **A** and **B** matrices may be easier in practice than stacking influence curves, M-estimation, if applicable, may be a more likely choice for variance estimation. Additionally, if the influence curve is known, one may define: $y(O_i; q_0) = l(q_0) \quad (q \quad q_0)$. The latter allows one to apply the M-estimator approach, even when $\hat{q}$ is not an M-estimator.

5.1.3 The Bootstrap

A widely applicable way to obtain statistical inference and 95% CIs is the bootstrap method [25]. Briefly, imagine we sample n units with replacement, drawn from our original sample of n units; we repeat the resampling process B number of times, so that we have B resulting samples, called bootstrap samples. Within each bootstrap sample, we apply the method we used to estimate the ATE from the original sample (e.g., regression standardization as described before), to obtain B separate ATE_i , with $i = 1; :::; B$ from each bootstrap sample. Then, the bootstrap estimate of the standard error shall be [25]:

$$SE_{Bootstrap} = \sqrt{\frac{1}{B \quad 1} \sum_{i=1}^{B} (ATE_i \quad \bar{ATE})^2} \tag{5.7}$$

where $\bar{ATE}$ is the mean across the $\hat{ATE}_i$ bootstrap sample specific estimates, i.e., $\bar{ATE} = \frac{1}{B} \sum_{i=1}^{B} ATE_i$. We then may apply the $SE_{Bootstrap}$ to obtain the corresponding bootstrap confidence intervals (CIs). For example, to construct the so-called normal-approximation bootstrap CI, we write:

$$\hat{ATE} \quad z_{1 \quad \frac{\alpha}{2}} SE_{Bootstrap}; \hat{ATE} + z_{1 \quad \frac{\alpha}{2}} SE_{Bootstrap} \tag{5.8}$$

where, $z_{1 \quad \frac{\alpha}{2}}$ is the $(1 \quad \frac{\alpha}{2})^{th}$ percentile of the standard normal distribution. Other methods to calculate the bootstrap confidence intervals are the percentile (for which, it is usually advisable to use a greater number of bootstrap samples), the bias corrected, and the bias corrected and accelerated methods, however, wherever we applied the bootstrap, calculation of the 95% was done by the normal based CI [25].

The validity of the bootstrap estimate for the standard error depends on two factors: (i) the number of the bootstrap replicates, which is finite, and, (ii) the validity of the so-called bootstrap principle, i.e., the distribution of the parameter estimate we target ($\hat{ATE}$ in our settings) given the ATE, is approximated by the distribution of estimates $\hat{ATE}$ given $\hat{ATE}$ - e.g., see [17].

5.2 Research Paper II

Below, we present a research paper published in Statistics in Medicine, where we have (i) offered a formula to calculate the 95% confidence intervals for different IPTW estimators that correspond to different target populations and (ii) provided a comprehensive step-by-step implementation in R software. In the main manuscript, the step-by-step derivation of the formulas is given via linearization, while in the corresponding appendix B, it is provided via M-estimation.

5.3 Extensions and further investigation

Via linearization, we obtained a formula that is general to both the PS weights (i.e., target population of question) and the model of choice for the treatment assignment mechanism. We have also shown that the same unifying formula can be obtained via M-estimation, when we apply the logit link to model the treatment assignment mechanism (Appendix B). A straightforward step would be to demonstrate the equivalence of the two formulas for any link function applied to the PS model - briefly, we could simply replace the row corresponding to the score equations of the *logit* link for the PS, with a general formulation for the link function of the PS.
Another following step could be to generalize these formulas for two-level IPTW estimators (e.g., for the marginal one (see eq. (4.5)), that uses a random effects model for the PS, it would be interesting to investigate whether it is feasible to adapt the score equation to correspond to that modeling approach). Of note, the model assumptions should always align with the required M-estimation assumptions, i.e., observations must be independent and the $y(:)$ function independent from i and n. In the setting of two-level structures, Y_i are not independent; they are correlated within the same cluster. That may indicate that our unit of focus should become the cluster - in that case, units will be independent in our setting.
Other possible extensions could be examining the performance of the closed-form variance estimators for the ATE, ATT, ATO (when taking into account the PS estimation step vs. when not), when the link function for the PS is different than the logit, as for our performed simulations (see 5.2), we focused on the logit link for the PS model. In the same manner, when the outcome is modeled via a different link than the logit (e.g., clog-log) or is of more complex nature than binary or categorical, such as time-to-event [36], it would be interesting to examine the comparative performance of the approximate variance estimators and evaluate them under different simulation scenarios.

A few points to highlight

We need to clarify the use of the term "robust variance estimator", and to which estimators it refers: although the two analytical (sandwich) estimators can be described as robust, whether or not the PS estimate is taken into account depends on whether or not the score equation corresponding to the PS model is included when the estimating equations are stacked. In the literature, there is often a

misconception as to what the term robust refers. There are certain cases, where the bootstrap does not provide valid results in variance estimation for PS methods. For example, in [8] they refer to the terms "naïve" and "robust" variance estimators, for the former, meaning a model-based estimator, which does not take into account neither the correlation between weighted observed outcomes nor the estimation of the PS for the construction of the weights. For the latter, they refer to an estimator that considers the correlation between weighted outcomes, but not the estimation of the PS.
Regarding the contribution to existing tools to estimate the variance of IPTW estimators, such as the bootstrap, it is reassuring that the three compared methods coincide in results, as they validate each other. For example, it is proven that the bootstrap no longer provides valid results in the case of matching on the PS; e.g., see [2, 9] and discussion section in [8], where they write: "...When using propensity-score matching with replacement, Abadie and Imbens found that the use of bootstrapping to estimate standard errors was inappropriate...However, when matching on the propensity score without replacement, Austin and Small found that bootstrapping performed well...".

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	lsh2005314	**Title**	Ms
First Name(s)	Andriana		
Surname/Family Name	Kostouraki		
Thesis Title	Comparison of causal inference methods for observational data with a hierarchical structure		
Primary Supervisor	Aurelien Belot		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Statistics in Medicine		
When was the work published?	15 April 2024		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I was the lead author of this paper. I suggested the basic concept which was further developed and shaped by collaboration with the Co-authors. I did the data analysis and drafted the manuscript. Co-authors performed the simulations included, which I critically revised. They also critically revised the manuscript, contributed to the analytical derivations presented in the paper and offered helpful insight and feedback on the manuscript.

SECTION E

Student Signature	[redacted]
Date	16/07/2024

Supervisor Signature	[redacted]
Date	16/07/2024

Received: 17 March 2023 | Revised: 12 February 2024 | Accepted: 1 April 2024

DOI: 10.1002/sim.10078

TUTORIAL IN BIOSTATISTICS

Statistics in Medicine WILEY

On variance estimation of the inverse probability-of-treatment weighting estimator: A tutorial for different types of propensity score weights

Andriana Kostouraki[1] | **David Hajage**[2] | **Bernard Rachet**[1] | **Elizabeth J. Williamson**[3] | **Guillaume Chauvet**[4] | **Aurélien Belot**[1] | **Clémence Leyrat**[3]

[1]Inequalities in Cancer Outcomes Network, Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

[2]Département de Santé Publique, Centre de Pharmacoépidémiologie (Cephepi), CIC-1901, Sorbonne Université, Inserm, Institut Pierre-Louis d'Epidémiologie et de Santé Publique, AP-HP, Hôpital Pitié-Salpêtrière, Paris, France

[3]Department of Medical Statistics, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

[4]ENSAI, CNRS, IRMAR-UMR 6625, Rennes University, Rennes, France

Correspondence
Andriana Kostouraki, Inequalities in Cancer Outcomes Network, Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK.
Email: andriana.kostouraki@lshtm.ac.uk

Funding information
Cancer Research UK, Grant/Award Numbers: C7923/A29018, C7923/A30945; Medical Research Council, Grant/Award Number: MR/T032448/1

Propensity score methods, such as inverse probability-of-treatment weighting (IPTW), have been increasingly used for covariate balancing in both observational studies and randomized trials, allowing the control of both systematic and chance imbalances. Approaches using IPTW are based on two steps: (i) estimation of the individual propensity scores (PS), and (ii) estimation of the treatment effect by applying PS weights. Thus, a variance estimator that accounts for both steps is crucial for correct inference. Using a variance estimator which ignores the first step leads to overestimated variance when the estimand is the average treatment effect (ATE), and to under or overestimated estimates when targeting the average treatment effect on the treated (ATT). In this article, we emphasize the importance of using an IPTW variance estimator that correctly considers the uncertainty in PS estimation. We present a comprehensive tutorial to obtain unbiased variance estimates, by proposing and applying a unifying formula for different types of PS weights (ATE, ATT, matching and overlap weights). This can be derived either via the linearization approach or M-estimation. Extensive R code is provided along with the corresponding large-sample theory. We perform simulation studies to illustrate the behavior of the estimators under different treatment and outcome prevalences and demonstrate appropriate behavior of the analytical variance estimator. We also use a reproducible analysis of observational lung cancer data as an illustrative example, estimating the effect of receiving a PET-CT scan on the receipt of surgery.

KEYWORDS
ATE, ATT, IPTW, matching weights, overlap weights, variance estimator

Aurélien Belot and Clémence Leyrat contributed equally to this study.

1 | INTRODUCTION

Propensity score (PS) methods were originally introduced by Rosenbaum and Rubin for the estimation of the causal effect of a treatment (or exposure) on an outcome in observational studies prone to confounding.[1] The PS for an individual is defined as the probability of receiving treatment conditional on a set of measured pre-treatment covariates. Once the PS is estimated (under the assumptions of consistency, positivity, no unmeasured confounding and no model misspecification), it can be used to balance individuals' characteristics between treatment groups and therefore remove confounding bias. Four different uses of the PS have been described in the literature: stratification or subclassification on the propensity score, adjustment on the propensity score, matching on the propensity score and inverse-probability-of-treatment weighting (IPTW).[1-5] After applying one of these methods, patients are on average expected to have similar characteristics between the different treatment groups. Thus, unlike multivariable outcome regression followed by some sort of standardization as in G-computation, PS approaches[5,6] focus on modeling the treatment allocation process instead of the outcome. This is one of the assets of propensity scores, in cases of rare outcomes or when the specification of the outcome model is very complex.

Among PS methods, IPTW has gained popularity and is widely implemented in observational studies and more recently, also in randomized controlled trials (RCTs). The main difference between the two is that in observational studies, the imbalance on patient baseline characteristics is most of the time systematic, whereas in RCTs a chance imbalance is usually the case. In RCTs, IPTW always increases precision in the effect estimate.[7,8] Briefly, in randomized settings there is no need for inclusion of variables that are predictive of the treatment (as treatment is randomized by design). Hence, one may include all the variables predictive of the outcome in the analysis—these are proven to reduce the variance while not biasing the effect estimate.[9] In this way, using PS modeling in the randomized setting always adds to the precision of the point estimate. IPTW has the advantage of targeting the same marginal estimand, no matter what covariates are adjusted for. We focus on IPTW particularly because it has more attractive mathematical properties. Propensity score matching often does not make use of the full sample and is harder to implement when the estimand of interest is the average treatment effect (ATE). On the other hand, stratification leads to more biased effect estimates when compared to various weighted estimators.[10] Lastly, covariate adjustment using the PS does not separate the design from the analysis stage.[11]

IPTW easily accommodates different measures of associations (eg, risk difference, marginal risk ratio and marginal odds ratio) and is widely used to estimate the ATE or the average treatment effect in the treated (ATT). Nonetheless, a problem frequently encountered is large weights, which leads to very imprecise effect estimates. There are a few solutions to this: (1) trimming the tails of the PS distribution, that is, remove individuals who have values of the PS that fall outside a specified range of the PS distribution—this however shifts the target population to a new one, which is no longer clearly defined,[12,13] (2) truncating the weights,[14] that is, set the value of weights that are greater than percentile p to the value of percentile p, or (3) using an estimand that focuses on a population for which extreme weights do not occur. The first two are simple and easy, and therefore popular. The third, while more complex, has an explicitly defined interpretation. Two popular recent versions of (3) are overlap (OW) and matching weights (MW).[15-17]

PS methods can be viewed as two-step estimators: in the first step we estimate the PS value for each individual. In the second step, the point estimate of the treatment effect and its corresponding standard error (SE) are obtained (eg, by applying a weighted regression model and a robust *sandwich* variance estimator). This robust estimator usually takes into account the dependence in the data introduced by the weighting procedure, while it may or *may not* correct for the PS estimation. In practice unfortunately, the robust estimator applied[18] incorrectly assumes that the individual propensity scores estimated from the first step were quantities known with certainty. Hereafter, we refer to this estimator as the *uncorrected estimator*. When targeting the ATE, this always results in an erroneously inflated variance and conservative confidence intervals.[19] When looking at the ATT, not accounting for a source of variability may lead to either over or underestimation of the variance, as shown by Reifeis and Hudgens.[20] Using an uncorrected variance estimator could lead to confidence intervals which are invalid, in the sense that we may end up with a type I error rate that is not equal to the advertised one. Non-parametric bootstrap can be used to obtain valid SEs, but the bootstrap may be computationally intensive for large databases, which motivated the development of analytic approaches.

Several authors have proposed closed-form estimators which provide unbiased IPTW variance estimates, mostly focusing on the ATE weights. This article is mainly driven by the work of Williamson et al in the context of randomized

trials,[8] as well as by the results of Lunceford and Davidian.[10] Williamson et al suggested an analytic variance estimator for the traditional ATE weights, when the risk difference, the marginal risk ratio and the marginal odds ratio is the treatment effect of interest. We extend this formula to a unifying variance estimator for different PS weighting schemes (ATE and ATT weights, OW and MW). We illustrate the case of a binary treatment and a binary outcome, but the formulas are generalizable to continuous outcomes as well. Published papers rarely use the correct way to estimate the variance, and one reason may be the absence of a step-by-step guide to do so. Hence we aim to: (i) offer a unifying formula to calculate the 95% confidence intervals that correspond to the IPTW estimates, applicable to different types of PS weights and (ii) provide a comprehensive step-by-step implementation in R software. In this way, we hope to contribute in disseminating the statistically reliable approach to estimate the SEs when using IPTW, as obtaining point estimates is half the inferential problem.[21] The article is organized as follows: Section 2 presents a brief introduction to the potential outcomes framework and the assumptions under which causal effects can be identifiable. Section 3 briefly describes the lung cancer study that will be used as an illustrative example in Section 5 . Section 4 offers an intuition for the derived formula for variance estimation for different PS weighting schemes, via the linearization method. Section 5 demonstrates the analytic calculation of the marginal large-sample IPTW variance estimator for different types of weights. Section 6 presents a Monte Carlo simulation to investigate the behavior of the estimators under different treatment and outcome prevalence values. Finally, we close in Section 7 with a discussion.

2 | CAUSAL FRAMEWORK AND GENERAL ASSUMPTIONS

We describe briefly how the IPTW estimator is derived and under which assumptions this is an unbiased estimator of the causal effect of a binary treatment on an outcome in observational studies.

2.1 | Notation, definition and assumptions

Consider an observational study of $i = 1, \dots, n$ individuals. We denote the observed individual outcome as Y_i (where Y_i, binary or continuous), the observed treatment as Z_i (1: treatment, 0: no treatment) and a set of measured baseline characteristics acting as potential confounders, $\boldsymbol{X}$—where, $\boldsymbol{X}$ an $n \times (p + 1)$ design matrix, with p the total number of covariates. Herein, vectors or matrices are presented in **bold** and by default a vector defines a column vector.

Now, assume Y_{1i} the potential outcome of individual i if they were to receive treatment level 1 and correspondingly Y_{0i} the potential outcome of individual i if they were to receive treatment level 0. Unfortunately, it is not feasible to observe both of these two outcomes simultaneously, leaving only one of these being observed for each individual (with the other being counterfactual). We can describe the treatment effect by contrasts of $\mu_1 = \mathrm{E}[Y_1]$ and $\mu_0 = \mathrm{E}[Y_0]$, such as risk differences. For example, for the ATE, we would have $\mu_1 - \mu_0 = \mathrm{E}[Y_1] - \mathrm{E}[Y_0]$. In RCTs, when targeting the overall population, we can write: $\mathrm{E}[Y|Z = 1] - \mathrm{E}[Y|Z = 0] = E[Y_1|Z = 1] - \mathrm{E}[Y_0|Z = 0] = \mathrm{E}[Y_1] - \mathrm{E}[Y_0]$, thanks to randomization. Contrary to that, in observational studies treatment is not randomly allocated between treatment groups, creating systematic imbalances of individual characteristics—thus, the last equality no longer holds. If these imbalances are not accounted for in the analysis, the treatment effect estimates will be biased. To remedy this, we define a propensity score estimator, which is named the inverse-probability-of-treatment weighted estimator (IPTW) and will estimate the causal treatment effect, in two steps:

1. Estimation of the PS.
2. Derivation of the estimates for the estimand of interest (eg, ATE).

For the causal effect of a treatment on an outcome to be identifiable, the below assumptions must be met:

(i) consistency[22];
(ii) positivity[23];
(iii) no unmeasured confounding (or conditional exchangeability).[24]

The above assumptions are strong and untestable (however, positivity and no unmeasured confounding are valid by design in RCTs). Causal consistency involves both the assumptions of (i) treatment variation irrelevance, and (ii) no interference.[25] We note that identification, although required, is not sufficient for estimation.[26,27] Hence, these assumptions are essential so that the causal interpretation holds, along with correct specification of the PS model. The former is vital for identifiability, whereas the latter is needed for the estimation step.

2.2 | Inverse probability-of-treatment weighted estimator of treatment effect

First, we define the propensity score e_i for individual i, as the probability of receiving treatment Z_i conditional on pre-treatment characteristics X_i. Suppose that $Z_i|X_i$ follows a generalized linear model:

$$e_i \equiv e(x_i) = Pr(Z_i = 1|X_i = x_i) = g^{-1}(\boldsymbol{x}_i^{\square}\square), \tag{1}$$

where g is a specific link function. In our demonstration, we derive the formulas for any possible link function. The most widely applied regression model is logistic regression, which uses the logit link function: $g(x) = \ln\left(\frac{x}{1-x}\right)$, but other link functions may be applied as well (see Section 4). Data adaptive methods are also available to estimate the PS, however this falls outside the scope of this article.[28-30] The PS model should include at least all confounders, and, potentially, predictors of the outcome. Once the chosen model is estimated, the estimated PS could be obtained for each individual:

$$\hat{e}_i = g^{-1}(\boldsymbol{x}_i^{\square}\hat{\square}), \tag{2}$$

where $\mathbf{x}_i$ is the $p + 1$ column vector of pre-treatment covariate values and $\hat{\square}$ is the estimated coefficients vector (including the intercept). For example, for logistic regression: $\hat{e}_i = \frac{\exp(\mathbf{x}_i^{\square}\hat{\square})}{1+\exp(\mathbf{x}_i^{\square}\hat{\square})}$.

Second, we can estimate the population means under treatment and under no treatment, that is, $\square_1 = \frac{E[Y_1 f(x)]}{E[f(x)]}$ and $\square_0 = \frac{E[Y_0 f(x)]}{E[f(x)]}$, with the target population depending on $f(x)$. When function $f(x)$ is equal to 1, we target the mean of the full population. When equal to the PS, we target the mean of the population of the treated. Finally, for values equal to $e(x)(1 - e(x))$ or to $\min(e(x), 1 - e(x))$, we target the overlap population via the overlap and matching weights respectively. The choice of the target population determines the type of weights, w_{1i}, w_{0i} leading to the resulting estimators, motivated by the Hájek estimator[31]:

$$\hat{\square}_1 = \left(\sum_{i=1}^{\square} Y_i Z_i \hat{w}_{1i}\right)\left(\sum_{i=1}^{\square} Z_i \hat{w}_{1i}\right)^{-1} \tag{3}$$

and

$$\hat{\square}_0 = \left(\sum_{i=1}^{\square} Y_i (1 - Z_i) \hat{w}_{0i}\right)\left(\sum_{i=1}^{\square} (1 - Z_i) \hat{w}_{0i}\right)^{-1}. \tag{4}$$

How to choose the appropriate weights is the topic of the next section.

2.3 | Types of propensity score weights

Usually, the research question defines the target population—that is, the group of individuals for whom inference is to be made. The estimand—the true value that one targets to estimate—is in turn determined by the target population (see Table 1). In this setting, weighting achieves two things: (i) re-balances individuals in the data so that those less represented in one treatment group are given more weight to compensate; and (ii) re-balances characteristics overall to reflect the target population. Consequently, depending on the target population, we choose the respective type of weight. We now discuss what populations are considered by the different weighting schemes: ATE weights, ATT weights, MW and OW.[32]

2.3.1 | ATE weights

To estimate the average effect of a treatment on an outcome if the whole population were to receive treatment vs if the whole population were not to, we apply the corresponding ATE weights (see Table 1). The ATE is useful when one is to implement a treatment/policy and so forth on the whole population and is interested in observing the effect of that treatment on the whole population. For the ATE to be identifiable, the assumption of conditional exchangeability need to hold for the whole population—a rather strong assumption to be made. Positivity needs to hold for both the whole population and the sample (two observed treatment groups). In practice, it is often the latter that is usually the problem—this however, may be relaxed by the following weighting schemes.

2.3.2 | ATT weights

For the ATE on the outcome if the population of the treated were to be treated vs if the population of the treated were not to be treated, we define the ATT weights. An important setting in which the ATT is applied is when investigating the effect of withholding a harmful treatment from a population with characteristics similar to those who are being treated (eg, the effect of preventing smoking from pregnant women on the birthweight of infants—in this setting, there is no interest in estimating the effect of withholding smoking from every pregnant woman vs not withholding it; the public health question surrounds the impact of preventing smoking among those who currently *do* smoke). For the ATT, assumptions of positivity and conditional exchangeability are relaxed (conditional exchangeability need to hold for the potential outcomes under no treatment only, while positivity requires patients to have non-zero propensity not to receive any treatment). In the same manner, we don't need everyone to have a non-zero propensity to receive treatment.

2.3.3 | Matching (MW) and overlap (OW) weights

In some cases, extreme weights happen due to positivity violations and/or strong confounding effects, which lead to biased estimates and excessive variance.[13,33,34] These are more likely when targeting the ATE rather than the ATT. Extreme weights might be an indication that a *new* population is of interest—this is because extreme weights occur when certain types of people are very likely to receive one treatment, thus are not necessarily amenable to intervention.

This new population of interest can be defined as those who achieve the "most overlap in observed characteristics between treatment groups," to cite Li et al.[15] In this situation, one targets the ATE in the overlap population or often described as the *equipoise* population (ATO).[32] The ATO can be easily estimated by applying either the so-called overlap weights (OW) or the matching weights (MW)—with the latter being a weighting analogue to 1:1 caliper matching without replacement or *pair matching*.[16]

The ATO answers questions regarding the subset of patients that usually have approximately equal chances of either receiving the treatment or not. This is a population very similar to the patients who would be enrolled in a clinical trial.

TABLE 1 Types of propensity score weights.

Target population	Estimand	Weight	Treated, $Z = 1$, $\hat{w}_{1i}$[a]	Untreated, $Z = 0$, $\hat{w}_{0i}$
Overall	ATE[b]	ATE	$\frac{1}{\hat{e}_i}$	$\frac{1}{1-\hat{e}_i}$
Treated	ATT[c]	ATT	1	$\frac{\hat{e}_i}{1-\hat{e}_i}$
Overlap or clinical equipoise	ATO[d]	MW[e]	$\frac{\min(\hat{e}_i,1-\hat{e}_i)}{\hat{e}_i}$	$\frac{\min(\hat{e}_i,1-\hat{e}_i)}{1-\hat{e}_i}$
Overlap or clinical equipoise	ATO	OW[f]	$1-\hat{e}_i$	$\hat{e}_i$

[a]Estimated weight for individual *i*.
[b]Average treatment effect.
[c]Average treatment effect in the treated.
[d]Average treatment effect in the overlap.
[e]Matching weights.
[f]Overlap weights.

In practice, application of the ATO is typically a consequence of having used MW or OW, rather than a conscious decision to target inference at the equipoise population. Despite that, it is important to have in mind the correspondence between weighting scheme of choice and population of interest. Regarding the identifiability assumptions, both the conditional exchangeability and the positivity assumptions are weaker than for the ATE—as these need to hold *only* within the overlap population.

Weighting techniques to target the ATO compared to those for the ATE or ATT are relatively recent. However, ATO methods have a growing body of use. This enhances the importance of having a correct variance estimator at hand that can be applied universally to any of the aforementioned settings.

3 | ILLUSTRATIVE EXAMPLE: NON-SMALL CELL LUNG CANCER (NSCLC) DATA

To illustrate the use of the IPTW variance estimator for various weighting schemes, we analyse a population-based dataset of patients diagnosed in England with a non-small cell lung cancer (NSCLC). Details on the main aims of this study can be found on the original article.[35] For the purposes of this tutorial, we will estimate the effect of PET-CT scan receipt on the probability of receiving surgery when targeting (i) the overall population of patients (ATE), (ii) those who actually receive PET-CT scan (ATT), and (iii) those with equal chances of either receiving a PET-CT or not (ATO).

The ATE answers "how the probability of receiving surgery would have been affected if all NSCLC patients (of stage I to III) at cancer diagnosis had received a PET-CT scan," a practice that has not been implemented yet everywhere in the UK health system. PET-CT scan is usually recommended to cancer patients before undertaking surgery, as it offers reliable knowledge about the localization and the spread of the tumor. Therefore, the treatment of interest is a binary variable which indicates if a PET-CT scan was undertaken, while the outcome is surgery with curative intent (no: 0; yes: 1; n = 10 398 NSCLC patients diagnosed in England in 2012; 6898 received a PET-CT scan). Both risk differences or risk ratios are appropriate measures for causal interpretations. When it comes to odds ratios, the fact that they are not collapsible quantities may make the comparison of results from different models harder; also, it is more difficult to interpret (because of the concept of odds). We emphasize that subsequent analyses are mostly for educational purposes, thus we did not account for other complexities in the data that otherwise must be accounted for such as missing values (we conducted complete-case analyses) or competing risk of death.

Now, we consider the validity of the required assumptions. The potential confounders accounted for are: sex, age at diagnosis, deprivation score, performance status, and stage of cancer at the time of cancer diagnosis; these have been extracted from an assumed causal diagram (Figure 1)—however, these may not be a sufficient set to consider (eg, variables like type of hospital could be considered as well). Regarding consistency and no-interference, we assume those to be

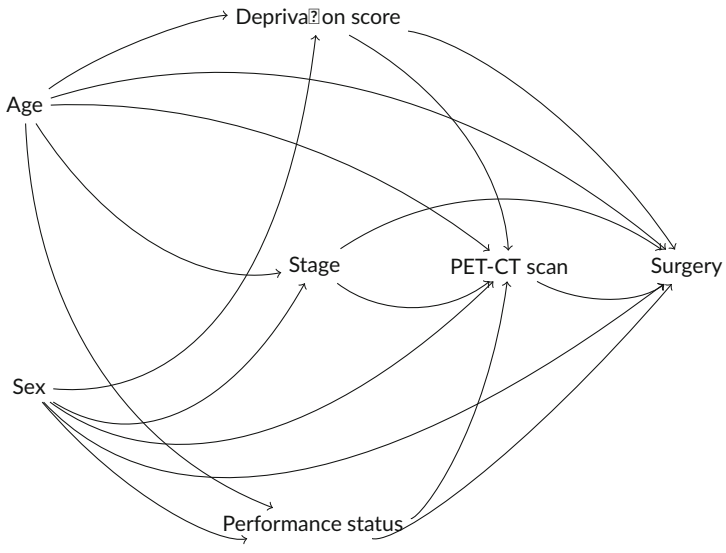


FIGURE 1 Assumed directed acyclic graph (DAG): Pre-treatment covariates acting as potential confounders.

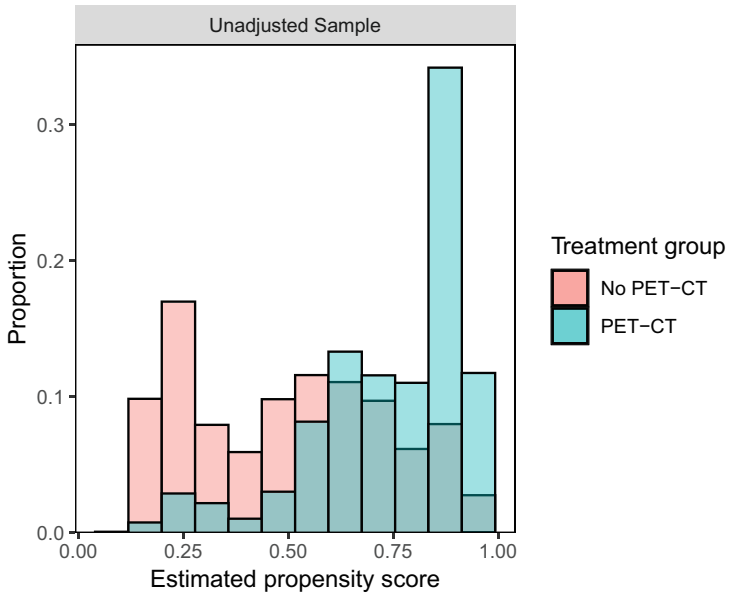


FIGURE 2 Histograms of the estimated propensity scores (*before weighting*) among those not having (pink) and having (cyan) received a PET-CT scan in the overall population, $n = 10\,398$.

TABLE 2 Characteristics of $n = 10\,398$ patients diagnosed with non-small cell lung cancer in England in 2012 by PET-CT scan receipt.

	No PET-CT scan	PET-CT scan	SMD[a]
n	3500	6898	-
Sex[b] = female (%)	1621 (46.3)	3091 (44.8)	0.030
Age at diagnosis (mean (SD))	75.58 (10.56)	70.46 (9.54)	0.509
Deprivation score (mean (SD))	0.17 (0.11)	0.17 (0.12)	0.023
Performance status:[c] poor (%)	2045 (58.4)	1369 (19.8)	0.861
Stage II of cancer at diagnosis:[d] yes (%)	461 (13.2)	1601 (23.2)	0.262
Stage III of cancer at diagnosis:[e] yes (%)	2340 (66.9)	2471 (35.8)	0.653

[a]Standardized mean difference.
[b]Reference category: male.
[c]Reference category: good.
[d]Reference category: no.
[e]Reference category: no.

valid in practice. Nonetheless, when empirically examining the distributions of the propensity scores within those having received a PET-CT scan vs those having not, we notice sufficient overlap between the two treatment groups, especially for PS values that are greater than 0.5, for the chosen PS model (see Figure 2, where the PS model included the main effects of age, sex, deprivation score, stage and performance status, as presented in Table 2).

4 | DERIVATION OF INVERSE PROBABILITY-OF-TREATMENT WEIGHTING VARIANCE ESTIMATOR VIA LINEARIZATION

4.1 | Causal effect measures of interest

Williamson et al[8] applied M-estimation to derive a variance estimator that accounts for the two IPTW estimation steps (see Section 2.1). M-estimation is an estimating method that can be applied for variance estimation of a target parameter,

among many other settings.[36,37] For more details, the reader may consult Sections 1 and 5 in Appendix I. In the following subsections we present an alternative approach, based on linearization—a more intuitive method and applicable to estimate the variance of a target parameter, under broad assumptions.[38] Section 1.5 of Appendix I demonstrates that the formulas for variance estimation derived from linearization coincide with those derived from M-estimation, when applying the logit link to estimate the PS. To begin with, the parameter of interest ψ is a function of the two marginal means (ie, $\psi = g(\mu_1, \mu_0)$). We shall denote as:

$$\hat{\psi}_1 = \hat{\mu}_1 - \hat{\mu}_0, \tag{5}$$

the marginal risk difference estimator (or mean difference estimator if the outcome is quantitative rather than binary),

$$\hat{\psi}_2 = \log(\hat{\mu}_1/\hat{\mu}_0), \tag{6}$$

the log marginal risk ratio estimator, and,

$$\hat{\psi}_3 = \log([\hat{\mu}_1/(1-\hat{\mu}_1)]/[\hat{\mu}_0/(1-\hat{\mu}_0)]), \tag{7}$$

the log marginal odds ratio estimator.

For the last two, researchers typically target the marginal risk ratio and marginal odds ratio. Nonetheless, we estimate the natural logarithms of these quantities, as these are more likely to be asymptotically normally distributed, making the construction of 95% confidence intervals a rather straightforward procedure. One may then exponentiate to approximate the 95% confidence intervals of the actual quantities rather than their logarithms.

In the following subsection, we present the main principles of linearization of a parameter estimator. Then, we derive each element used in the final variance estimator from each of the estimation stages described in 2.1: (1) the propensity model and the estimated weights, and (2) the population means under treatment or under no treatment, and the marginal treatment effect. Details on intermediate steps of the final derivations via linearization not shown in the main manuscript are presented in Appendix I.

4.2 | Introduction to linearization

Linearization is a method applied to estimate the variance of an estimator $\hat{\theta}$ of a parameter θ. The parameter may be the population mean μ, or a treatment effect ψ.[38] In linearization, we seek for an artificial variable $l_i(\hat{\theta})$, which we shall call the *linearized variable of* $\hat{\theta}$, to express the difference between the estimator $\hat{\theta}$ and the parameter θ, as a sum of $l_i(\hat{\theta})$, plus a negligible quantity, $o_p(n^{-1/2})$—which approaches zero as the sample size increases at a rate of $n^{-1/2}$, where n is the sample size:

$$\hat{\theta} - \theta = \frac{1}{n}\sum_{i=1}^{n} l_i(\hat{\theta}) + o_p(n^{-1/2}). \tag{8}$$

Equation (8) ensures that $Var(\hat{\theta}) \approx n^{-1}Var\{l(\hat{\theta})\}$ (note that in our setting, $l_i(\hat{\theta})$ are assumed *i.i.d*). Hereafter, a statistic or parameter estimator, $\hat{\theta}$, for which we can derive a linearized variable $l_i(\hat{\theta})$, shall be called *linearizable*.

In our setting, we aim to estimate the variance of the causal treatment effect estimator of interest (ie, $\hat{\psi}_1$ or $\hat{\psi}_2$ or $\hat{\psi}_3$). This ceases to be a straightforward task, as we need to account both for the correlation of the individual outcomes after weighting and the uncertainty in the PS weights estimation. This motivates to prove that the corresponding estimator is linearizable—that is, can be written in a form such as Equation (8). For example, for the variance of $\hat{\psi}_1$, we must identify $l_i(\hat{\psi}_1)$, such that:

$$\hat{\psi}_1 - \psi_1 = \frac{1}{n}\sum_{i=1}^{n} l_i(\hat{\psi}_1) + o_p(n^{-1/2}). \tag{9}$$

From Equation (5):

$$(\hat{\mu}_1 - \hat{\mu}_0) - (\mu_1 - \mu_0) = \frac{1}{n}\sum_{i=1}^{n} l_i(\hat{\mu}_1 - \hat{\mu}_0) + o_p(n^{-1/2}). \tag{10}$$

By rearranging terms in (10):

$$(\hat{\theta}_1 - \theta_1) - (\hat{\theta}_0 - \theta_0) = \frac{1}{n}\sum_{i=1}^{n} I_i(\hat{\theta}_1 - \hat{\theta}_0) + o_p(n^{-1/2}) \Leftrightarrow \tag{11}$$

$$(\hat{\theta}_1 - \theta_1) - (\hat{\theta}_0 - \theta_0) = \frac{1}{n}\sum_{i=1}^{n} \{I_i(\hat{\theta}_1) - I_i(\hat{\theta}_0)\} + o_p(n^{-1/2}) \Leftrightarrow \tag{12}$$

$$(\hat{\theta}_1 - \theta_1) - (\hat{\theta}_0 - \theta_0) = \frac{1}{n}\sum_{i=1}^{n} I_i(\hat{\theta}_1) - \frac{1}{n}\sum_{i=1}^{n} I_i(\hat{\theta}_0) + o_p(n^{-1/2}) \Leftrightarrow \tag{13}$$

$$(\hat{\theta}_1 - \theta_1) - (\hat{\theta}_0 - \theta_0) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} I_i(\hat{\theta}_1) + o_p(n^{-1/2})}_{I_i(\hat{\theta}_1)\text{-assumed linearized variable for } \hat{\theta}_1} - \{\frac{1}{n}\sum_{i=1}^{n} I_i(\hat{\theta}_0) + o_p(n^{-1/2})\}, \tag{14}$$

where, to derive line 12 from 11, we apply the so-called linearization rules presented in Reference 38; specifically, the second rule, states that the linearized statistic of a sum of estimators is equal to the sum of the linearized statistics of the estimators. As a result, to estimate the variance of $\hat{\theta}_1$, one needs to derive a linearized statistic for each of the two population mean estimators, θ_1 and θ_0. For the variances of $\hat{\theta}_2$ and $\hat{\theta}_3$, since these two are not linear functions of the estimators for the population means, the first rule of linearization is applied, which states that the linearized statistic of a derivable function, f of a vector of estimators, is the product of the linearized statistics of the estimators and the matrix of the partial derivatives of the function f. More on this is demonstrated in Section 4.5. Now, to linearize $\hat{\theta}_1$, we may write:

$$\hat{\theta}_1 - \theta_1 = \frac{1}{n}\sum_{i=1}^{n} I_i(\hat{\theta}_1) + o_p(n^{-1/2}) \Leftrightarrow \tag{15}$$

$$\left(\sum_{i=1}^{n} Y_i Z_i \hat{w}_{1i}\right)\left(\sum_{i=1}^{n} Z_i \hat{w}_{1i}\right)^{-1} - \theta_1 = \frac{1}{n}\sum_{i=1}^{n} I_i(\hat{\theta}_1) + o_p(n^{-1/2}). \tag{16}$$

And, likewise for $\hat{\theta}_0$:

$$\hat{\theta}_0 - \theta_0 = \frac{1}{n}\sum_{i=1}^{n} I_i(\hat{\theta}_0) + o_p(n^{-1/2}) \Leftrightarrow \tag{17}$$

$$\left(\sum_{i=1}^{n} Y_i(1 - Z_i)\hat{w}_{0i}\right)\left(\sum_{i=1}^{n} (1 - Z_i)\hat{w}_{0i}\right)^{-1} - \theta_0 = \frac{1}{n}\sum_{i=1}^{n} I_i(\hat{\theta}_0) + o_p(n^{-1/2}). \tag{18}$$

From expression (16), we see that $\hat{\theta}_1$ is a function of Y_i, Z_i and $\hat{w}_{1i}$. The first two, are observed quantities, while the third component is a parameter estimated from the sample. Hence, to linearize the population mean under treatment, we must linearize $\hat{w}_{1i}$. The same thought process goes for the linearization of $\hat{\theta}_0$. However, both of $\hat{w}_{1i}$ and $\hat{w}_{0i}$ are functions of the estimated PS (see Table 1). Consequently, we must start by the linearization of the PS model and the respective weights, before moving to linearize the population means. This is presented in the following two subsections.

Of note, $I_i(\hat{\theta})$ cannot be applied directly for variance estimation, since $I_i(\hat{\theta})$ usually depends on unknown parameters. Replacing these unknown quantities with their estimators leads to the *estimated linearized variable* $\hat{I}_i(\hat{\theta})$, and to the variance estimator[38]:

$$\hat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{n}\hat{\mathrm{Var}}(\hat{I}(\hat{\theta})) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left\{\hat{I}_i(\hat{\theta}) - \frac{1}{n}\sum_{j=1}^{n}\hat{I}_j(\hat{\theta})\right\}^2. \tag{19}$$

Lastly, we mention that linearization is similar to the delta method and most of the time leads to identical variance estimators. In the following subsections, we exploit the fact that Taylor expansion may be applicable to our settings—which simplifies greatly the variance estimation procedure. A brief explanation of Taylor expansion and the delta method, can be found in the Glossary section of Appendix I.

4.3 | Linearization of the propensity model and propensity score weights

Assume that $Z_i|X_i$ follows the generalized linear model of Equation (1) and that the estimator $\hat{\beta}$ is obtained from solving the generalized estimating equation:

$$\bar{U}(\beta) \equiv \frac{1}{n}\sum_{i=1}^{n} U_i(\beta) = 0, \tag{20}$$

with

$$\begin{aligned} U_i(\beta) &\equiv \left(\frac{\partial g^{-1}(\boldsymbol{x}_i^{\top}\beta)}{\partial \beta}\right)^{\top} \times \mathrm{Var}(Z_i = 1|X_i = \boldsymbol{x}_i)^{-1} \times (Z_i - g^{-1}(\boldsymbol{x}_i^{\top}\beta)) \\ &= \frac{1}{g'(g^{-1}(\boldsymbol{x}_i^{\top}\beta))} \times \boldsymbol{x}_i \times \frac{1}{g^{-1}(\boldsymbol{x}_i^{\top}\beta)(1 - g^{-1}(\boldsymbol{x}_i^{\top}\beta))} \times (Z_i - g^{-1}(\boldsymbol{x}_i^{\top}\beta)) \\ &= \frac{(Z_i - g^{-1}(\boldsymbol{x}_i^{\top}\beta))\boldsymbol{x}_i}{g^{-1}(\boldsymbol{x}_i^{\top}\beta)(1 - g^{-1}(\boldsymbol{x}_i^{\top}\beta))g'(g^{-1}(\boldsymbol{x}_i^{\top}\beta))} \\ &= \frac{(Z_i - e_i)\boldsymbol{x}_i}{e_i(1 - e_i)g'(e_i)}, \end{aligned} \tag{21}$$

and with $g'(\cdot)$, the first derivative of the link function, $g(\cdot)$. We note that the first line in Equation (21) is the score equation for β and can be directly obtained from the generalized estimating equations theory.[39] Estimators $\hat{\beta}$ in Equation (20) could be alternatively seen as M-estimators.[37] We have $\bar{U}(\hat{\beta}) - \bar{U}(\beta) = -\bar{U}(\beta)$, and under mild conditions:

$$\bar{U}(\hat{\beta}) - \bar{U}(\beta) = E\left[\frac{\partial U(\beta)}{\partial \beta}\right](\hat{\beta} - \beta) + o_p(n^{-1/2}). \tag{22}$$

More details on the derivation of Equation (22) can be found in Appendix I.

We have:

$$E\left[\frac{\partial U(\beta)}{\partial \beta}\right] = E\left[\frac{\partial\{\frac{(Z-e)x}{e(1-e)g'(e)}\}}{\partial \beta}\right] = -E\left[\frac{\boldsymbol{x}\boldsymbol{x}^{\top}}{e(1-e)\{g'(e)\}^2}\right]. \tag{23}$$

From Equations (21)–(23), we obtain:

$$\begin{aligned} &\hat{\beta} - \beta = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{C}^{-1}\frac{(Z_i - e_i)\boldsymbol{x}_i}{e_i(1 - e_i)g'(e_i)} + o_p(n^{-1/2}), \\ &\text{with } \boldsymbol{C} = E\left[\frac{\boldsymbol{x}\boldsymbol{x}^{\top}}{e(1-e)\{g'(e)\}^2}\right]. \end{aligned} \tag{24}$$

This general formulation allows logistic regression to be considered as one method among others for estimating the propensity score, and we provide several examples of generalized linear models. Under a logistic model, we have

$$g(x) = \ln\left(\frac{x}{1-x}\right) \text{ and } g'(x) = \frac{1}{x(1-x)}, \tag{25}$$

and Equation (24) can be rewritten as:

$$\hat{\theta} - \theta = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{C}^{-1}(Z_i - e_i)\boldsymbol{x}_i + o_p(n^{-1/2}),$$
$$\text{with } \boldsymbol{C} = E\left[e(1-e)\boldsymbol{x}\boldsymbol{x}^{\top}\right]. \tag{26}$$

Under a probit model, we have

$$g(x) = \Phi^{-1}(x) \text{ and } g'(x) = \frac{1}{\phi(x)}, \tag{27}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ the cumulative density and probability density functions of a standard normal distribution, respectively. Equation (24) can be rewritten as

$$\hat{\theta} - \theta = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{C}^{-1}\frac{(Z_i - e_i)\phi\left(\boldsymbol{x}_i^{\top}\theta\right)}{e_i(1-e_i)}\boldsymbol{x}_i + o_p(n^{-1/2}),$$
$$\text{with } \boldsymbol{C} = E\left[\frac{\phi\{\left(\boldsymbol{x}^{\top}\theta\right)\}^2\boldsymbol{x}\boldsymbol{x}^{\top}}{e(1-e)}\right]. \tag{28}$$

Under a complementary log-log model (cloglog), we have

$$g(x) = \ln(-\ln(1-x)) \text{ and } g'(x) = -\frac{1}{(1-x)\ln(1-x)}, \tag{29}$$

and Equation (24) can be rewritten as

$$\hat{\theta} - \theta = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{C}^{-1}\frac{(Z_i - e_i)\ln(1-e_i)\boldsymbol{x}_i}{e_i} + o_p(n^{-1/2}),$$
$$\text{with } \boldsymbol{C} = E\left[\frac{1-e}{e}\{\ln(1-e)\}^2\boldsymbol{x}\boldsymbol{x}^{\top}\right]. \tag{30}$$

Similar developments can be obtained for other modeling techniques, provided that an explicit estimating equation is available (which excludes tree-based regression approaches, for example, generalized boosted models[29]).

Whatever the underlying regression model, Equation (24) gives a linear approximation of $\hat{\theta} - \theta$, which captures the uncertainty in the estimation of the propensity model coefficients. Estimating the variance of these coefficients is not the objective pursued here, but this approximation leads to one of the components in the linearized variable for the propensity score weight, and to the first element of the linearized variable $l_i(\hat{\theta})$. For this purpose, we first obtain a Taylor expansion for the estimated weights. We denote as $\hat{w}_{1i} \equiv w_{1i}(\hat{\theta})$ and $\hat{w}_{0i} \equiv w_{0i}(\hat{\theta})$:

$$\hat{w}_{1i} - w_{1i} = w'_{1i}(\hat{\theta} - \theta) + o_p(n^{-1/2}),$$
$$\text{with } w'_{1i} = \frac{\partial w_{1i}}{\partial\theta^{\top}}(\theta), \tag{31}$$

and

$$\hat{w}_{0i} - w_{0i} = w'_{0i}(\hat{\theta} - \theta) + o_p(n^{-1/2}),$$
$$\text{with } w'_{0i} = \frac{\partial w_{0i}}{\partial\theta^{\top}}(\theta). \tag{32}$$

Therefore, for each possible weighting scheme, we need the corresponding derivative. For the weights summarized in Table 1, these derivatives are functions of $\boldsymbol{x}_i$ and e_i, and are shown in Table 3.

TABLE 3 Derivatives of weight functions.

Target population	Estimand	Weight	Treated, $Z = 1, \boldsymbol{w}_{1i}^{'}$		Untreated, $Z = 0, \boldsymbol{w}_{0i}^{'}$	
Overall	ATE[a]	ATE	$-\frac{\boldsymbol{x}_i^T}{e_i^2 g^{'}(e_i)}$		$\frac{\boldsymbol{x}_i^T}{(1-e_i)^2 g^{'}(e_i)}$	
Treated	ATT[b]	ATT	0		$\frac{\boldsymbol{x}_i^T}{(1-e_i)^2 g^{'}(e_i)}$	
Overlap or clinical equipoise	ATO[c]	MW[d]	0	if $e_i < 0.5$	$\frac{\boldsymbol{x}_i^T}{(1-e_i)^2 g^{'}(e_i)}$	if $e_i < 0.5$
			$-\frac{\boldsymbol{x}_i^T}{e_i^2 g^{'}(e_i)}$	if $e_i > 0.5$	0	if $e_i > 0.5$
Overlap or clinical equipoise	ATO	OW[e]	$-\frac{\boldsymbol{x}_i^T}{g^{'}(e_i)}$		$\frac{\boldsymbol{x}_i^T}{g^{'}(e_i)}$	

[a]Average treatment effect.
[b]Average treatment effect in the treated.
[c]Average treatment effect in the overlap.
[d]Matching weights. Note that matching weights have a non-differentiable point at 0.5; see "A weighting analogue to pair matching in propensity score analysis" and "Comments on 'A weighting analogue to pair matching in propensity score analysis' by L. Li and T. Greene."
[e]Overlap weights.

4.4 | Linearization of the population means

From the definition of $\hat{\mu}_1$ in Equation (3), we have:

$$\hat{\mu}_1 - \mu_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{w}_{1i} Z_i (Y_i - \mu_1)}{E[w_1 Z]} + o_p(n^{-1/2}). \tag{33}$$

By plugging (26) and (31) into (33), we obtain (after some algebra):

$$\begin{aligned} \hat{\mu}_1 - \mu_1 &= \frac{1}{n}\sum_{i=1}^{n} l_i(\mu_1) + o_p(n^{-1/2}), \\ \text{with } l_i(\mu_1) &= \frac{w_{1i} Z_i (Y_i - \mu_1)}{E[w_1 Z]} + \boldsymbol{D}\frac{\{Z_i - e_i\}\boldsymbol{x}_i}{e_i(1-e_i)g^{'}(e_i)} \\ \text{and } \boldsymbol{D} &= \frac{1}{E[w_1 Z]} E[w_1^{'} Z (Y - \mu_1)] \times \boldsymbol{C}^{-1}, \end{aligned} \tag{34}$$

where $\boldsymbol{C}$ is given in Equation (24). The linearized variable $l_i(\mu_1)$ contains two components. The first one is the usual linearized component associated with the estimating equation for μ_1, and incorporates the corresponding variability, while the second one incorporates the variability associated with the estimation of γ.

The estimated linearized variable of $\hat{\mu}_1$ is obtained by replacing into (34) all the unknown quantities by estimators. Finally, we obtain the estimated linearized variable:

$$\hat{l}_i(\mu_1) = l_{1i}^{\text{uncor}} + l_{1i}^{\text{cor}} \tag{35}$$

with

$$l_{1i}^{\text{uncor}} = \frac{1}{n^{-1}\sum_{j=1}^{n} \hat{w}_{1j} Z_j} \{\hat{w}_{1i} Z_i (Y_i - \hat{\mu}_1)\} \tag{36}$$

and

$$l_{1i}^{\text{cor}} = \frac{1}{n^{-1}\sum_{j=1}^{n} \hat{w}_{1j} Z_j} \left\{ \frac{1}{n}\sum_{j=1}^{n} \hat{w}_{1j}^{'} Z_j (Y_j - \hat{\mu}_1) \right\} \left\{ \frac{1}{n}\sum_{j=1}^{n} \frac{\boldsymbol{x}_j \boldsymbol{x}_j^{\square}}{\hat{e}_j(1-\hat{e}_j)\{g^{'}(\hat{e}_j)\}^2} \right\}^{-1} \frac{(Z_i - \hat{e}_i)\boldsymbol{x}_i}{\hat{e}_i(1-\hat{e}_i)g^{'}(\hat{e}_i)}.$$

By using a similar proof, we obtain the estimated linearized variable for $\hat{\theta}_0$:

$$\hat{l}_i(\hat{\theta}_0) = l_{0i}^{\text{uncor}} + l_{0i}^{\text{cor}} \tag{37}$$

with

$$l_{0i}^{\text{uncor}} = \frac{1}{n^{-1}\sum_{j=1}^{n}\hat{w}_{0j}(1-Z_j)}\{\hat{w}_{0i}(1-Z_i)(Y_i-\hat{\theta}_0)\} \tag{38}$$

and

$$l_{0i}^{\text{cor}} = \frac{1}{n^{-1}\sum_{j=1}^{n}\hat{w}_{0j}(1-Z_j)} \left\{\frac{1}{n}\sum_{j=1}\hat{w}_{0j}'(1-Z_j)(Y_j-\hat{\theta}_0)\right\} \left\{\frac{1}{n}\sum_{j=1}\frac{\boldsymbol{x}_j\boldsymbol{x}_j^{\top}}{\hat{e}_j(1-\hat{e}_j)\{g'(\hat{e}_j)\}^2}\right\}^{-1} \frac{(Z_i-\hat{e}_i)\boldsymbol{x}_i}{\hat{e}_i(1-\hat{e}_i)g'(\hat{e}_i)}.$$

As mentioned in the introductory paragraph of this section, these two estimated linearized variables can be applied to estimate the variance of the estimated population means: $\hat{\text{Var}}(\hat{\theta}_1) = \frac{1}{n}\hat{\text{Var}}(\hat{l}_i(\hat{\theta}_1))$ and $\hat{\text{Var}}(\hat{\theta}_0) = \frac{1}{n}\hat{\text{Var}}(\hat{l}_i(\hat{\theta}_0))$. The estimated linearized variables have also the advantage of explicitly highlighting the different components used to calculate the counterfactual means $\hat{\theta}_1$ and $\hat{\theta}_0$. If we limit the estimated linearized variables to their first term in Equations (35) and (37), that is, l_{1i}^{uncor} and l_{0i}^{uncor} respectively, then $\frac{1}{n}\hat{\text{Var}}(\hat{l}_i(\hat{\theta}_1))$ and $\frac{1}{n}\hat{\text{Var}}(\hat{l}_i(\hat{\theta}_0))$ would estimate the variance of $\hat{\theta}_1$ and $\hat{\theta}_0$ as if the individual propensity score values were *known*; this corresponds to the *uncorrected* variance estimator. This estimator directly depends on the type of weights used. The uncorrected variance estimates are equal to the sandwich robust variance estimates, when this estimator does not account for the PS estimation. This estimator, accounts for the correlation between individual outcomes—as by weighting, a new, pseudo-population is generated, where the same individual may appear more than once when their weight is greater than 1. However, it deals with the propensity score predictions as if these were the true propensities of treatment assignment, rather than estimates. An equivalent formula of this, can be derived via M-estimation (see Appendix I).

On the other hand, including the second term, l_{1i}^{cor} in Equation (35) or l_{0i}^{cor} in (37), allows to obtain the *corrected* variance estimator, which additionally takes into account the fact that the propensity score values (and respective weights) are *estimated* from the predictions of a regression model. This can be broken down into two parts: the first (in gray) depends directly on the derivative of the chosen weighting scheme, while the second depends on the regression model used to estimate the propensity score. Therefore unlike some other estimators previously proposed in the literature, this expression supports the generalization of the variance estimator to any type of weighting scheme and to a wider variety of PS models.

4.5 | Linearization of causal treatment effect

Finally, we derive the estimated linearized variables for the estimators of the causal treatment effect measures. For the estimator of the risk (or mean for quantitative outcome) difference $\hat{\theta}_1$, we have:

$$\hat{l}_i(\hat{\theta}_1) = \hat{l}_i(\hat{\theta}_1) - \hat{l}_i(\hat{\theta}_0). \tag{39}$$

And, the respective estimated variance would be obtained from Equation (19):

$$\hat{\text{Var}}(\hat{\theta}_1) = \frac{1}{n}\hat{\text{Var}}(\hat{l}(\hat{\theta}_1)) = \frac{1}{n(n-1)}\sum_{i=1}\left\{\hat{l}_i(\hat{\theta}_1) - \frac{1}{n}\sum_{j=1}\hat{l}_j(\hat{\theta}_1)\right\}^2. \tag{40}$$

By using the computation rules for linearization given in Deville,[38] we obtain for the estimator of the marginal risk ratio $\hat{\theta}_2$ the estimated linearized variable:

$$\hat{l}_i(\hat{\theta}_2) = \frac{\hat{l}_i(\hat{\theta}_1)}{\hat{\theta}_1} - \frac{\hat{l}_i(\hat{\theta}_0)}{\hat{\theta}_0}, \tag{41}$$

and for the estimator of the marginal odds ratio $\hat{\theta}_3$ the estimated linearized variable:

$$\hat{I}_i(\hat{\theta}_3) = \frac{\hat{I}_i(\theta_1)}{\theta_1(1-\theta_1)} - \frac{\hat{I}_i(\theta_0)}{\theta_0(1-\theta_0)}, \tag{42}$$

And, correspondingly, their variance estimators:

$$\widehat{\text{Var}}(\hat{\theta}_2) = \frac{1}{n}\widehat{\text{Var}}(\hat{I}(\hat{\theta}_2)) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left\{\hat{I}_i(\hat{\theta}_2) - \frac{1}{n}\sum_{j=1}^{n}\hat{I}_j(\hat{\theta}_2)\right\}^2, \tag{43}$$

and,

$$\widehat{\text{Var}}(\hat{\theta}_3) = \frac{1}{n}\widehat{\text{Var}}(\hat{I}(\hat{\theta}_3)) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left\{\hat{I}_i(\hat{\theta}_3) - \frac{1}{n}\sum_{j=1}^{n}\hat{I}_j(\hat{\theta}_3)\right\}^2. \tag{44}$$

In the next section we provide the corresponding R code to generate the variance estimates. For the sake of brevity, the code focuses on the risk difference and overlap weights (targeting the ATO), with a propensity model estimated using a logistic regression. We used non-stabilized weights throughout, which for the case of a binary treatment at a single time-point should provide the same results to the stabilized ones.[26] Nevertheless, stabilized weights are not expected to provide the same results for cases of time-varying or continuous treatments. All the rest calculations of different weights and effect measures follow a similar manner, and can be found in Appendices I and II (corresponding to the theoretical derivations and R code application, respectively). Results for all the different combinations of weights and effect measures for the illustrative example are also provided in the following section.

5 | GUIDED IMPLEMENTATION

In Sections 5.1–5.3 we show a step by step implementation of the corrected variance estimator using R version 4.0.3. In Section 5.4, we present the results of applying this to the NSCLC data set.

We maintain the notation introduced in Section 2.1. The covariates are denoted as $X1$, $X2$, $X3$, $X4$, $X5$, $X6$ (following the order presented in Section 3) throughout the implementation. A description of the characteristics of patients diagnosed with NSCLC in England in 2012 stratified by receipt of PET-CT scan is presented in Table 2.

5.1 | Estimation of marginal means

First step: Prediction of the propensity scores $\hat{e}_i$.

The first required step of IPTW estimation is to fit the propensity score model.[26] Individual propensity scores are typically estimated using the predictions of a logistic regression model of Z on X. The predicted probabilities of receiving a PET-CT scan conditional on the measured confounders (Equation 2) are obtained in the box below:

Box 1 - estimated propensity scores

```
mydata <- readRDS("□/mydata.rds") # □ path to wd
ps_mod <- glm(Z ~ X1 + X2 + X3 + X4 + X5 + X6, family = binomial(), data = mydata) # fit the PS model (only
    logistic here)
mydata$e <- predict(ps_mod, type = "response") # predicted propensity scores
```

Following the propensity score estimation we estimate the weights (see Table 1) and weight derivatives (see Table 3). We then apply each weight type to target the respective marginal mean. Below we present the marginal means under treated and untreated, when targeting the overlap population:

Second step: Estimation of marginal means.

Box 2 - estimated marginal means

```
mydata$OW <- with(mydata, ifelse(Z == 1, 1-e, e)) # Overlap weights
## Weight derivatives estimation
X <- cbind(1, as.matrix(mydata[, c("X1", "X2", "X3", "X4", "X5", "X6")]))
OWp <- X*with(mydata, ifelse(Z == 1, -e*(1-e), e*(1-e)))
## Estimated marginal means
n1_OW <- sum(mydata$OW*mydata$Z)
mu1_OW <- sum(mydata$OW*mydata$Z*mydata$Y)/n1_OW
n0_OW <- sum(mydata$OW*(1 - mydata$Z))
mu0_OW <- sum(mydata$OW*(1 - mydata$Z)*mydata$Y)/n0_OW
```

We note that both the estimated PS and the weights can be alternatively automatically obtained via the R package WeightIt.[40]

5.2 | Estimated linearized variables: Uncorrected

Third step: Estimation of the uncorrected linearized variables.

In a third step, one may estimate the linearized variables for the marginal means under treatment and control (see formulas (35) and (37), respectively), while dismissing the second terms of Equations (35) and (37). This corresponds to the uncorrected linearized variables for the marginal means, the variance of which, divided by the sample size, n, is equal to the uncorrected variance estimate for $\hat{\Delta}_1$ and $\hat{\Delta}_0$. As stated in Section 4.2, the uncorrected variance estimator does not account for PS estimation.

Below, we present the uncorrected variance estimate calculation for the overlap weights and the risk difference.

Box 3 - estimated linearized variables - uncorrected

```
mydata$l1_OW_un <- with(mydata, n/n1_OW*(OW*Z*(Y-mu1_OW))) ### OW
mydata$l0_OW_un <- with(mydata, n/n0_OW*(OW*(1-Z)*(Y-mu0_OW)))
mydata$l_RD_OW_un <- with(mydata, l1_OW_un - l0_OW_un)
var(mydata$l_RD_OW_un)/n
```

Note that an equivalent approach to calculate the IPTW estimators in Equations (5)–(7), would be to fit the weighted regression model specific to the outcome of interest. For example when targeting the log odds ratio, the uncorrected variance estimates may be obtained from the geeglm command of the geepack R library that is used for GEE models (by selecting independence for the working correlation matrix) or via the library sandwich. However, these commands calculate the variance without considering the estimation of the PS, which provides either larger or smaller variance estimates for the ATT or ATO weights and always larger estimates when targeting the ATE.

5.3 | Estimated linearized variables: Corrected

Fourth step: Estimation of the corrected linearized variables.

To get the corrected variance estimates, one must include the second part of the formulas (35) and (37), when calculating the linearized variables. After that, the corrected variance estimates can be obtained for the marginal means, just like before. The difference here is that this estimator considers the fact that the individual PS were estimated. The additional term included in the corrected linearized variables depends on both the weighting scheme and the regression model applied to estimate the PS—a logit model for our illustrative example. Below we present the corrected variance estimate calculation for the overlap weights and the risk difference.

Box 4 - estimated linearized variables - corrected

```
xx <- t(apply(X, 1, function(x) cbind(x) %*% t(cbind(x)))) ## x_ix_i^T
C <- matrix(colSums(with(mydata, e*(1-e)*xx))/n, ncol(X), ncol(X))
Cinv <- solve(C)
### OW
mydata$l1_OW_cor <- mydata$l1_OW_un + drop(n/n1_OW*with(mydata, rbind(colSums(
    OWp*Z*(Y-mu1_OW))/n) %*% apply((Z-e)*X, 1, function(x) Cinv %*% cbind(x))))
mydata$l0_OW_cor <- mydata$l0_OW_un + drop(n/n0_OW*with(mydata, rbind(colSums(
    OWp*(1-Z)*(Y-mu0_OW))/n) %*% apply((Z-e)*X, 1, function(x) Cinv %*% cbind(x))
    ))

mydata$l_RD_OW_cor <- with(mydata, l1_OW_cor - l0_OW_cor)
var(mydata$l_RD_OW_cor)/n
```

Box 5 - R packages PSweight and geex

```
#PSweight application
library(PSweight) #load PSweight library
mod_OW <- PSweight(ps.formula = Z  X1 + X2 + X3 + X4 + X5 + X6,data = mydata,weight = 'overlap', yname =
    'Y') #Overlap weights
summary(mod_OW, type = "DIF", CI = TRUE) # to be compared with sqrt(var(mydata$l_RD_OW_cor)/n)
#geex application
library(geex)
estfun <- function(data, model){
  L <- model.matrix(model, data = data)
  Z <- model.response(model.frame(model, data = data))
  Y <- data$Y
  function(theta){
    p <- length(theta);p1 <- length(coef(model));lp <- L %*% theta[1:p1];rho <- plogis(lp)
    w <- ifelse(Z==1, 1-rho, rho) #Overlap weights
    score_eqns <- apply(L, 2, function(x) sum((Z - rho)*x))
    mu1 <- w*(Z==1)*(Y - theta[p-1])
    mu0 <- w*(Z==0)*(Y - theta[p])
    c(score_eqns, mu1, mu0)
  }
}
res_OW <- m_estimate(estFUN = estfun, data = mydata, roots = c(coef(ps_mod), mu1_OW, mu0_OW),compute_roots
    = FALSE,
outer_args = list(model = ps_mod))
vcov(res_OW) #mod_OW$covmu
```

5.4 | Results

Standard errors obtained from the uncorrected, corrected large-sample variance and bootstrap estimator appear in Table 4 along with the IPTW estimates for the effect of PET-CT scan on the receipt of surgery and their corresponding 95% confidence intervals.

In this particular example, results show all uncorrected variance estimates to be larger than the analytic ones (Table 4). As anticipated, bootstrap estimates are almost identical to those analytically derived. Bootstrap was performed non-parametrically on the entire procedure (ie, including PS estimation for each separate bootstrap sample). We provide the normal-based 95% CI for 1000 bootstrap repetitions in Table 4. As for the point estimates, we interpret the risk difference when targeting the ATO: "had the population of those with equal chance of either receiving a PET-CT scan or not received a PET-CT scan, the probability of undertaking surgery would have increased by 19.3% compared to had the same population not received any." Lastly, across the different association measures we notice a difference between the ATE/ATT effect estimates vs the ATO.

6 | SIMULATION STUDY

6.1 | Aim

We performed a simulation study to assess the behavior of the *analytic correction* of the variance, as compared to the *uncorrected variance estimator* in finite samples. By doing so, we will be able to highlight scenarios where the variance

TABLE 4 Point estimates and standard errors (SEs) for the different effect measures of PET-CT scan on the probability of receiving surgery[a].

	Estimand	Weight	Estimate	Uncorrected SE[b]	Corrected SE[c]	Bootstrap SE[d]	Uncorrected 95% CI	Corrected 95% CI	Bootstrap normal-based 95% CI
Risk difference	ATE[e]	ATE	0.155	0.016	0.013	0.014	(0.123, 0.186)	(0.128, 0.181)	(0.128, 0.181)
	ATT[f]	ATT	0.156	0.020	0.017	0.018	(0.118, 0.195)	(0.122, 0.190)	(0.121, 0.191)
	ATO[g]	Overlap	0.193	0.010	0.008	0.008	(0.174, 0.212)	(0.177, 0.209)	(0.177, 0.209)
	ATO	Matching	0.194	0.009	0.008	0.008	(0.177, 0.211)	(0.179, 0.209)	(0.179, 0.209)
Log marginal risk ratio	ATE	ATE	0.470	0.060	0.050	0.051	(0.353, 0.588)	(0.371, 0.569)	(0.370, 0.571)
	ATT	ATT	0.377	0.056	0.050	0.051	(0.267, 0.487)	(0.279, 0.475)	(0.277, 0.477)
	ATO	Overlap	0.900	0.059	0.051	0.051	(0.785, 1.015)	(0.799, 1.001)	(0.800, 1.001)
	ATO	Matching	1.037	0.061	0.054	0.054	(0.918, 1.156)	(0.931, 1.143)	(0.930, 1.143)
Log marginal odds ratio	ATE	ATE	0.704	0.082	0.069	0.070	(0.543, 0.865)	(0.569, 0.839)	(0.567, 0.841)
	ATT	ATT	0.647	0.087	0.077	0.078	(0.478, 0.817)	(0.496, 0.798)	(0.494, 0.801)
	ATO	Overlap	1.152	0.070	0.060	0.060	(1.015, 1.289)	(1.034, 1.270)	(1.034, 1.270)
	ATO	Matching	1.281	0.070	0.062	0.062	(1.143, 1.420)	(1.160, 1.403)	(1.159, 1.404)

[a]For n = 10 398 patients diagnosed with non-small cell lung cancer in England in 2012; 6898 received a PET-CT scan.
[b]Does not account for the uncertainty in PS estimates.
[c]Accounts for the uncertainty in PS estimates.
[d]Accounts for the uncertainty in PS estimates; 1000 non-parametric bootstrap replications.
[e]Average treatment effect.
[f]Average treatment effect in the treated.
[g]Average treatment effect in the overlap.

may be either over or underestimated when using the uncorrected variance estimator, and illustrate the importance of applying a corrected variance estimator for correct inference.

6.2 | Data-generating mechanisms and estimands

We randomly generated 10 independent and identically distributed baseline characteristics $X_1, X_2, \ldots, X_{10} \sim N(0,1)$ for $n = 10\,000$ individuals. Values for the binary treatment variable Z were drawn from a Bernoulli(p_Z) for each individual (we suppressed the individual indicator i for brevity), with:

$$p_Z = \text{expit}\{\alpha_{0,Z} + \alpha_L(X_1 + X_2 + X_3) + \alpha_M(X_4 + X_5 + X_6) + \alpha_H(X_7 + X_8 + X_9) + \alpha_{VH}X_{10}\}, \tag{45}$$

where $\text{expit}(x) = \frac{\exp(x)}{1+\exp(x)}$, $\alpha_L = \log(1.1)$, $\alpha_M = \log(1.25)$, $\alpha_H = \log(1.5)$ and $\alpha_{VH} = \log(2)$ represent low, medium, high and very high effects of covariates on treatment. Each characteristic was assumed to have the same effect on outcome as it does on treatment. Therefore, a binary outcome was generated for each individual from a Bernoulli(p_Y), with:

$$p_Y = \text{expit}\{\beta_{0,Y} + \beta Z + \beta_L(X_1 + X_2 + X_3) + \beta_M(X_4 + X_5 + X_6) + \beta_H(X_7 + X_8 + X_9) + \beta_{VH}X_{10}\}, \tag{46}$$

In Equation (46), β denotes the conditional log(OR) relating the treatment Z to the outcome Y. Lastly, $\beta_{0,Z}$, $\beta_{0,Y}$ and β were set to values that induce the desired treatment prevalence π_Z, event rate π_Y and marginal effect Γ (RD, RR, or OR in combination with ATE, ATT, MW, or OW estimator) in the simulated sample. So, for each paired combination of effect measure (RD, RR, OR) and estimand (ATE, ATT, MW, OW), there is a different set of parameter values. The process to identify these parameter values used a minimization approach, which is described in detail in Appendix I.

We allowed the following parameters to vary across simulations:

1. Treatment prevalence: $\pi_Z \in \{0.10, 0.25, 0.50\}$;
2. Event rate: $\pi_Y \in \{0.10, 0.25, 0.50\}$;
3. Marginal treatment effect (ATE, ATT, MW, ATO). Five increasing treatment effects were evaluated for each measure, their values depending on the type of measure:
   a. risk difference: $\Gamma_{RD} \in \{-0.20, -0.10, 0, 0.10, 0.20\}$;
   b. logarithm of relative risk: $\Gamma_{\text{RR}} \in \{\log(1/1.50), \log(1/1.20), \log(1), \log(1.20), \log(1.50)\}$;
   c. logarithm of odds ratio: $\Gamma_{\text{OR}} \in \{\log(1/2.25), \log(1/1.50), \log(1), \log(1.50), \log(2.25)\}$;

Hence, for each paired combination of effect measure and estimand we applied a full factorial design with a total of $3 \times 3 \times 5 = 45$ data-generating mechanisms.

6.3 | Methods of analysis

For the analysis methods we applied the true propensity score models and for the variance of the estimated treatment effect we compare the performance when (i) we do not account for the PS estimation step (uncorrected variance estimator) and (ii) the PS estimation step is accounted for (corrected variance estimator).

6.4 | Performance measures

For each scenario, we used $n_{\text{sim}} = 10\,000$ replicates to calculate the following performance criteria[41]):

1. Bias: $\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \hat{\Gamma}_i - \Gamma$;
2. Coverage: proportion of times Γ is enclosed in the confidence interval, calculated as $\hat{\Gamma}_i \pm Z_{1-\frac{\alpha}{2}} \times \hat{\text{SE}}(\hat{\Gamma}_i)$, where $1-\alpha$ is the confidence level and $Z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution;

3. Relative % error in average model-based standard error (ModSE): $100 \times \left(\frac{\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \hat{\text{SE}}(\hat{\Gamma}_i)}{\sqrt{\frac{1}{n_{\text{sim}}-1} \sum_{i=1}^{n_{\text{sim}}} (\hat{\Gamma}_i - \bar{\Gamma})^2}} - 1 \right)$, where $\hat{\text{SE}}(\hat{\Gamma}_i) = \sqrt{\hat{V}(\hat{\Gamma}_i)}$ is the estimated standard error of treatment effect $\hat{\Gamma}_i$ (obtained using the corrected or uncorrected estimator).

Relative (%) error in ModSE allows the evaluation of the performance of the variance estimators: a value > 0 (or < 0) suggests that model-based standard errors overestimate (respectively, underestimate) the variability of the treatment effect estimate. Regarding coverage, it evaluates if the procedure for constructing the confidence interval achieves the advertised nominal level (95% here).

6.5 | Simulation results

Bias is reported separately in Appendix I in Figures 1–3 for the risk difference, logarithm of the risk ratio and logarithm of the odds ratio estimates, respectively; this was nearly zero across scenarios, as anticipated. When using the corrected variance estimator, model-based standard error was always very close to the empirical standard error; the relative (%) error in average model-based standard error was always very close to zero throughout scenarios (see Figures 6–8 in Appendix I). Hereafter we report coverage rate of the resulting 95% confidence intervals for the logarithm of the marginal risk ratio (for the risk difference and the logarithm of the marginal odds ratio, see Figures 4 and 5, respectively, in Appendix I).

On the left-hand side of Figure 3, we present the coverage values for the uncorrected variance estimator, whereas the right-hand side displays those for the corrected variance estimator. Looking at the log risk ratio (Figure 3), overcoverage for the uncorrected estimator was prevalent in the majority of the examined scenarios, while the corrected estimator achieved the advertised nominal level of 95%. A similar pattern was observed for both the risk difference and log odds ratio. However, for an assumed log risk ratio equal to 1.50, treatment prevalence of 10% and event rate of 50%, we observed undercoverage when targeting (i) the ATT and (ii) the ATO (either via overlap or matching weights). For the remaining scenarios, over-estimation was observed when using the uncorrected variance estimator. Relative (%) model-based standard error results align with those observed for the coverage results and are not shown here.

7 | DISCUSSION

We have provided a unifying formula for the corrected variance estimator, which treats IPTW weights as a special case of a wider class often referred to as *balancing weights*.[17] In addition, we have highlighted the risks of bias when applying the *uncorrected variance estimator* when the PS is modeled via a logistic model. The risks are especially prominent when one targets effects in the overall population, where overcoverage and conservative confidence intervals occur (Figure 3). We have shown via an illustrative example of examining the effect of PET-CT scan on receiving surgery in NSCLC patients that the corrected variance estimator may be applied for multiple estimands and effect measures via R software, either manually or via the packages PSweight and geex. In our simulations, we have illustrated that under scenarios of large-samples and very small treatment prevalence, the ATT and ATO variance estimates may be underestimated, which leads to incorrect inference.

When targeting the ATE, the difference between the corrected and uncorrected variance estimators (ie, accounting for PS estimation) consists of the negative of a quadratic form around a positive definite matrix (see Appendix I)[8]; therefore, the corrected variance estimate will always be lower than the uncorrected one. This result is not universal to different propensity score weighting schemes, as we have demonstrated in our simulations. For the rest, the correction term may be either positive or negative, respectively leading to either under or overestimation when omitted.

We note that the proposed unifying formulae (Equations 35 and 37) should be applicable to any type of propensity score weighting scheme as long as these are once differentiable functions of the propensity score. While matching weights are not differentiable at the value of 0.5, it can be proven that the matching weights estimator follows an asymptotically normal distribution, an assumption required for M-estimation to be applicable, but not necessary for linearization (see Appendix I and Reference 42). When applying techniques to ameliorate problems caused by extreme weights, like trimming, this formula is no longer valid, so use of weights that target the ATO are recommended as an alternative.

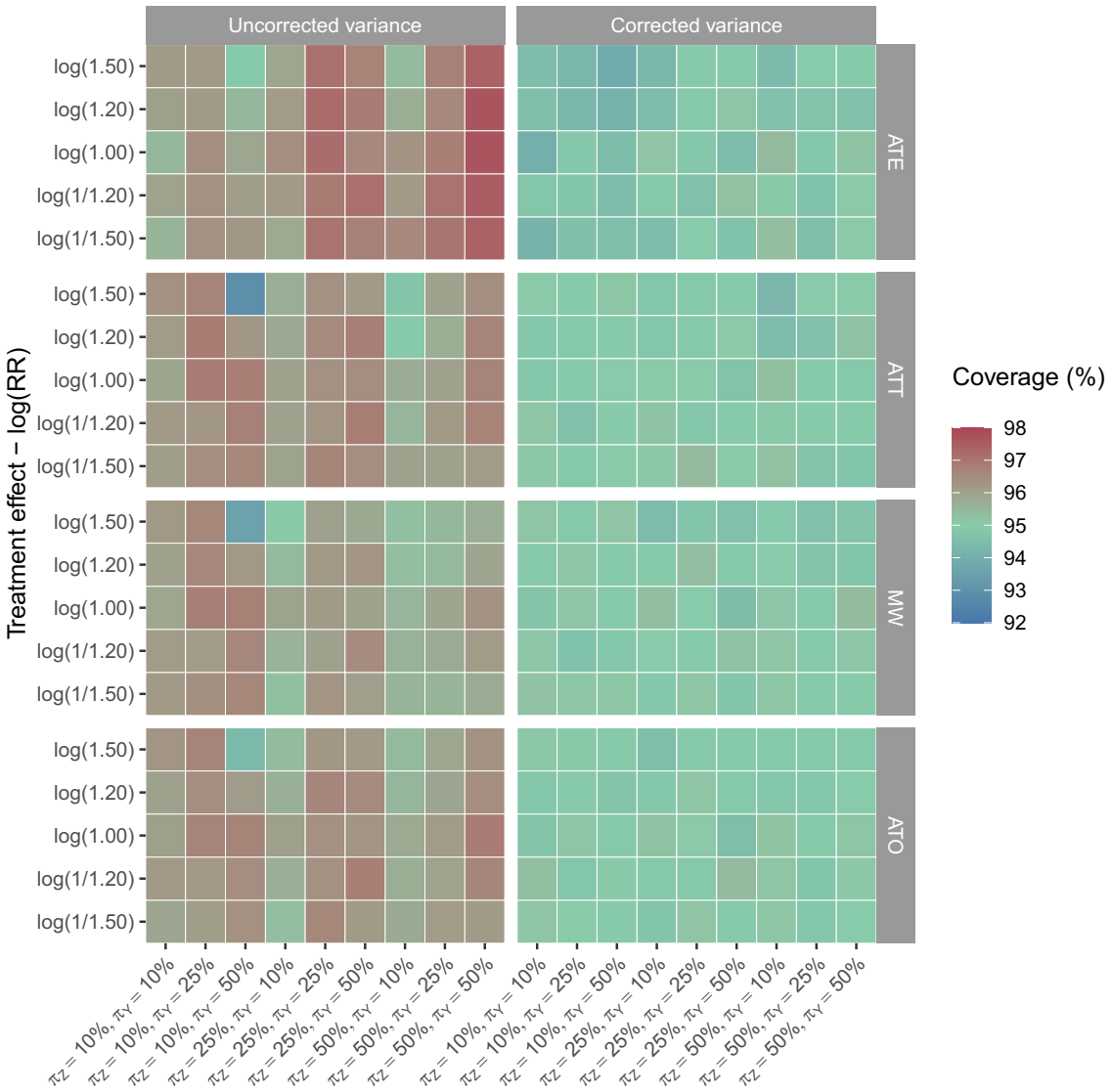


FIGURE 3 Coverage rate of 95% confidence intervals in simulation studies (n_{sim} = 10 000) comparing the corrected vs uncorrected variance estimator for the log risk ratio (RR) for four different propensity score (PS) weighting schemes: Average treatment effect (ATE), average treatment effect in the treated (ATT), matching weights (MW) and overlap weights (OW).

This work adds to the existing research around variance estimation when weighting by the estimated propensity score functions. Lunceford and Davidian[10] have applied M-estimation[37] to estimate the IPTW large-sample marginal variance of mean (or risk) differences for continuous (or binary) outcomes. Furthermore, Williamson et al have expanded the closed-form variance estimator for ATE weights for the natural logarithm of the marginal risk ratio and marginal odds ratio.[8] Lastly, Hajage et al, motivated by Austin,[43] have proposed an estimator for the marginal hazards ratio along with an R package for the implementation.[44]

The resulting formula (see Equation (11) in Appendix I), is an extension to Williamson et al's and coincides with the ATE-specific derivation. To our knowledge, calculations may be performed via the R packages *geex*[45] or *PSweight*,[46] as shown in Section 5. Briefly, *geex* automates the procedure of variance estimation via M-estimation, given the user stacks the estimating equations required to estimate the corresponding quantity of interest (eg, for the case of any weighting scheme, the estimating equation for the logistic PS model will have the same general form, etc.). This package can utilize any procedure that applies M-estimation, hence may be used within a much wider scope than that of variance estimation for IPTW. *PSweight* also offers a wide range of options to the user, as it supports a variety of balancing weights, multiple treatments, augmented weighting estimators and so forth.[46] It also allows for application of data-adaptive methods to estimate the PS, such as generalized boosted models (GBM).[29] However, our tutorial adds to the current functionality of

PSweight, as the former allows for application of any link function, compared to the latter, which provides only the case of the logit link. At the same time, the linearization method does not require the specific estimating equations—as *geex* does—which are not so simple with non-canonical links, such as probit.

In our illustrative example we used the non-parametric bootstrap[47] as the "gold standard" approach for the estimation of the variance. Bootstrap SEs were almost identical to the corrected ones across estimands and effect measures as expected, with the sample size being sufficiently large. However, a recent study by Austin,[48] comparing the performance of the bootstrap vs the corrected asymptotic variance estimator, showed that the bootstrap estimator gave more accurate estimates when sample size was ≤1000 under several treatment prevalences for the ATE and when the treatment prevalence was moderate to high for the ATT. Nonetheless, for the overlap and matching weights both methods under comparison performed equally well.

Comparing the corrected analytic estimator to the bootstrap was not the aim of this article, nonetheless it is important to mention that depending on the underlying scenario, the bootstrap may sometimes perform slightly better. The analytic estimator, similarly to the bootstrap, is asymptotic, which means that its performance becomes better as the sample size increases (recall, that in our simulations, the sample size chosen is 10 000, which allows the closed-form estimator to show its full potential). Despite that, the analytic estimator may be an alternative to the bootstrap under computationally demanding situations, such as when having very large databases. This last statement should need further research on "how large" is large enough for the closed-form estimator to have optimal performance (especially, under extreme scenarios of very low or very high treatment prevalences). If multiple imputation (MI) is used to handle missingness in the data, this raises problems combining the MI and the bootstrap.[49] Another situation that inflates both computational intensity and time, are numerical simulations, making the use of bootstrap within this framework less appealing. Lastly, ease of reproducibility is another aspect of the analytical estimator that makes it more attractive than the bootstrap, for which reproducibility requires not only the use of the exact same data set, but also the same seed number applied at the starting point of the bootstrap procedure.[44]

The formulas generated are based on large-sample theory, thus results are valid if the sample size of each treatment level is sufficiently large. However, even for point estimation, IPTW requires a large enough sample size anyway, so if a researcher is working with small samples, IPTW may not be the appropriate method to begin with. The uncorrected variance estimator is invalid when fitting a logistic regression model on the PS as demonstrated in our simulations. Of further research would be the evaluation of the performance of the corrected vs the uncorrected estimators across weighting schemes and different link functions for the PS. We have provided variance estimators for IPTW assuming the PS is modeled via a GLM; however, there are recently developed machine learning techniques, offered for PS prediction. Nonetheless for these, even the bootstrap may no longer be valid.[50,51] Moreover, modified versions for the IPTW estimators have been recently proposed. These modified smoothed versions suggest that they mitigate the large variance problem, which can be even greater in the presence of positivity violations, extreme weight values and so forth.[52] Additionally, another category of weights reported relatively recently[53] are the entropy weights. Although we have not applied the proposed formula to these, this should be straightforward. Lastly, regarding complex settings mentioned in section 3: for missing data problems (on confounders or on the outcome), IPTW can be used after multiple imputation. Assuming a correct specification of the imputation model, that must include the outcome, treatment effect estimates and variance estimates (applying the proposed variance estimator to account for the uncertainty in PS estimation, obtained from each imputed data set) can be combined to get an overall estimate via Rubin's rules.[54] For cases with competing risks, we must highlight that the target of inference is different and outside the scope of this study, however, an introduction to the relevant setting, estimands and variance estimation can be found in Reference 55.

To summarize, the proposed analytic estimator is a less computationally intensive approach than the bootstrap and with better statistical properties than the uncorrected estimator, with the latter likely to provide misleading inference. This tutorial was oriented for applied researchers and we provided many details on each step of the process for estimating the variance of the causal effect under study. Finally, we hope this work helps disseminate the use of the corrected variance estimator and improve the statistical inference when applying IPTW.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their constructive comments that significantly improved our manuscript. AK is supported by the Cancer Research UK Doctoral Studentship in Inequalities in Cancer Outcomes (CR UK Programme C7923/A30945); BR and AB are funded by the Cancer Research UK programme Grant (C7923/A29018); CL is supported by the UK Medical Research Council (Skills Development Fellowship MR/T032448/1).

DATA AVAILABILITY STATEMENT

The toy dataset used to run the example is provided as an .RData file in the supplementary material.

ORCID

Andriana Kostouraki https://orcid.org/0000-0003-4536-8987
David Hajage https://orcid.org/0000-0002-8475-4090
Elizabeth J. Williamson https://orcid.org/0000-0001-6905-876X
Aurélien Belot https://orcid.org/0000-0003-1410-5172

REFERENCES

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
2. Rosenbaum P, Rubin D. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*. 1984;79:516-524.
3. Rosenbaum P. Model-based direct adjustment. *J Am Stat Assoc*. 1987;82:387-394.
4. Austin P. An introduction to propensity-score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res*. 2011;46(3):399-424.
5. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from Naïve enthusiasm to intuitive understanding. *Stat Methods Med Res*. 2012;21(3):273-293.
6. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*. 1998;17(19):2265-2281.
7. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149-1156.
8. Williamson EJ, Forbes A, White IR. Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Stat Med*. 2014;33(5):721-737.
9. Westreich D, Cole SR, Funk MJ, Brookhart MA, Stürmer T. The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiol Drug Saf*. 2011;20(3):317-320.
10. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19):2937-2960.
11. Austin PC. Optimal Caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*. 2011;10(2):150-161.
12. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*. 2009;96(1):187-199.
13. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLoS One*. 2011;6(3):e18174.
14. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168(6):656-664.
15. Li F, Thomas LE, Li F. Addressing extreme propensity scores via the overlap weights. *Am J Epidemiol*. 2018;188(1):250-257.
16. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. *Int J Biostat*. 2013;9(2):215-234.
17. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. *J Am Stat Assoc*. 2018;113(521):390-400. doi:10.1080/01621459.2016.1260466
18. Freedman DA. On the so-called Huber sandwich estimator and robust standard errors. *Am Stat*. 2006;60(4):299-302.
19. Robins JM, Hernán M, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550-560.
20. Reifeis SA, Hudgens MG. On variance of the treatment effect in the treated when estimated by inverse probability weighting. *Am J Epidemiol*. 2022;191(6):1092-1097.
21. Webster-Clark M, Stürmer T, Wang T, et al. Using propensity scores to estimate effects of treatment initiation decisions: state of the science. *Stat Med*. 2020;40(7):1718-1735. doi:10.1002/sim.8866
22. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*. 2009;20(1):3-5.
23. Westreich D, Cole SR. Invited commentary: positivity in practice. *Am J Epidemiol*. 2010;171(6):674-677.
24. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol*. 1986;15(3):413-419.
25. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology*. 2009;20(6):880-883.
26. Hernán MA, Robins JM. *Causal Inference: What if*. London: Chapman and Hall; 2020.
27. Aronow PM, Robins JM, Saarinen T, Sävje F, Sekhon J. Nonparametric identification is not enough, but randomized controlled trials are. arXiv preprint arXiv:210811342, 2021.
28. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;29(3):337-346.
29. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med*. 2013;32(19):3388-3414. doi:10.1002/sim.5753
30. Smith MJ, Mansournia MA, Maringe C, et al. Introduction to computational causal inference using reproducible Stata, R, and Python code: a tutorial. *Stat Med*. 2021;41(2):407-432. doi:10.1002/sim.9234
31. Hajek J. Comment on "an essay on the logical foundations of survey sampling" by D. Basu. *Foundations of Statistical Inference*. New York: Holt, Rinehart, and Winston; 1971:236.

32. Greifer N, Stuart EA. Choosing the estimand when matching or weighting in observational studies. arXiv preprint arXiv:210610577, 2021.
33. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1-21.
34. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Serv Outcomes Res Methodol*. 2001;2(3-4):259-278.
35. Belot A, Fowler H, Njagi EN, et al. Association between age, deprivation and specific comorbid conditions and the receipt of major surgery in patients with non-small cell lung cancer in England: a population-based study. *Thorax*. 2019;74(1):51-59.
36. Boos DD, Stefanski LA. 7. *Essential Statistical Inference: Theory and Methods*. New York: Springer; 2013.
37. Stefanski LA, Boos DD. The calculus of M-estimation. *Am Stat*. 2002;56(1):29-38.
38. Deville JC. Variance estimation for complex statistics and estimators: linearization and residual techniques. *Surv Methodol*. 1999;25(2):193-203.
39. Hardin JW, Hilbe JM. *Generalized Estimating Equations*. London: Chapman and Hall/CRC; 2002.
40. Greifer N. WeightIt: weighting for covariate balance in observational studies. r package version 0.13.1; 2022. https://CRAN.R-project.org/package=WeightIt
41. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074-2102.
42. Orihara S, Kawamura T, Taguri M. Comments on 'a weighting analogue to pair matching in propensity score analysis' by L. Li and T. Greene. *Int J Biostat*. 2022;19:53-60.
43. Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Stat Med*. 2016;35(30):5642-5655.
44. Hajage D, Chauvet G, Belin L, Lafourcade A, Tubach F, De Rycke Y. Closed-form variance estimator for weighted propensity score estimators with survival outcome. *Biom J*. 2018;60(6):1151-1163.
45. Saul BC, Hudgens MG. The calculus of M-estimation in R with geex. *J Stat Softw*. 2020;92(2):1-15.
46. Zhou T, Tong G, Li F, Thomas L, Li F. PSweight: an R package for propensity score weighting analysis. arXiv preprint arXiv:201008893v4, 2021.
47. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. London: Chapman and Hall/CRC; 1993.
48. Austin PC. Bootstrap vs asymptotic variance estimation when using propensity score weighting with continuous and binary outcomes. *Stat Med*. 2022;41(22):4426-4443.
49. Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. *Stat Med*. 2018;37(14):2252-2266. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7654
50. Hines O, Dukes O, Diaz-Ordaz K, Vansteelandt S. Demystifying statistical learning based on efficient influence functions. *Am Stat*. 2022;76(3):292-304. doi:10.1080/00031305.2021.2021984
51. Gill RD. Non- and semi-parametric maximum likelihood estimators and the Von Mises method (part 1) [with discussion and reply]. *Scand J Stat*. 1989;16(2):97-128.
52. Liao J, Rohde C. Variance reduction in the inverse probability weighted estimators for the average treatment effect using the propensity score. *Biometrics*. 2021;78:660-667.
53. Matsouaka A, Roland LY, Zhou Y. Overlap, matching, or entropy weights: what are we weighting for? arXiv preprint arXiv:221012968, 2022.
54. Leyrat C, Seaman SR, White IR, et al. Propensity score analysis with partially observed covariates: how should multiple imputation be used? *Stat Methods Med Res*. 2019;28(1):3-19.
55. Young JG, Stensrud MJ, Tchetgen EJT, Hernán MA. A causal framework for classical statistical estimands in failure-time settings with competing events. *Stat Med*. 2020;39(8):1199-1236.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kostouraki A, Hajage D, Rachet B, et al. On variance estimation of the inverse probability-of-treatment weighting estimator: A tutorial for different types of propensity score weights. *Statistics in Medicine*. 2024;43(13):2672-2694. doi: 10.1002/sim.10078

Chapter 6

Discussion

In presence of hierarchical structures, as described in the previous chapters (i.e., nested, two-level structures or above), to account for clustering in our analyses plan should be a good idea - even as a form of sensitivity analyses compared to analyses that do not account for it. In the following sections, we summarize our main findings, present the key strengths and weaknesses of this work and close with matters of future research.

6.1 Summary of main findings

In this dissertation, we began by reviewing causal inference methods when applied within the single-level setting for observational data - namely, g-computation, IPTW and DR methods. We then continued by exploring potential subtleties that arise when we transit to the hierarchical (two-level) framework in terms of both the identification and definition of the target estimand. When treatment is assigned at the individual level and one targets a cs-ATE, positivity must be checked within clusters, which is a stronger assumption to make, compared to across-cluster positivity. When one targets a marginal ATE, in theory, positivity across clusters is required. In practice however, when applying a PS clustered estimator (like the non-parametric clustered estimator (4.7) we used in chapter 4), one must ensure within-cluster positivity holds for the sample, to make the estimation process possible, without having to resort to other solutions - like exclusion of clusters with treatment assigned at the cluster level, or grouping together clusters with similar characteristics to achieve having both treatment levels being assigned within the new groups. Therefore in a way, this type of estimators allows for straightforward checks of whether the targeted estimand is plausible to identify. This is contrast to the g-computation estimators applied in chapter 4, where, this is not made explicit - e.g., the estimator, in case of presence of clusters with only treatment or control being assigned, would normally pool information from a similar cluster - which should lead to some sort of extrapolation when interpreting the results. The assumption of no unmeasured confounding must hold both for individual- and cluster-level confounders. Alternatively, by conditioning on the cluster, elimination of

cluster-level confounding is automatically achieved. Populations such as those who actually receive treatment (*ATT*) or those who have equal chance of either receiving treatment or not (*ATO*) may be of interest in applications of cancer epidemiology. However, our focus in the hierarchical framework was the average treatment effect in the overall population (*ATE*), which is very frequently the target. In our investigations the main question was to what extent a misspecification of the outcome model would affect the $p \quad ATE$ estimates when applying mixed effects models to account for clustering under omission of a cluster-level confounder when standardization (g-computation) is applied. The same question (under misspecification of the PS model this time), was applied for PS weighted estimators, when these were combined with random effects models. The follow-up question was, whether mixed effects models enable inferring causal relationships in settings with serious concerns about unmeasured cluster-level confounding.

In our simulation studies (see section 4.4), we illustrated scenarios where the performance of g-computation estimators was better compared to their marginal PS weighting counterparts when combined with random effects (except for the non-parametric clustered IPTW estimator, which had the best performance across simulation scenarios), in terms of bias. This may be attributed to the fact that random effects do not provide balance on unmeasured cluster-level confounders; on the other hand, the objective of PS methods is to achieve balance on any measured confounders [54]. Another factor that must be taken into account, is the average sample size within cluster in our simulation design, which was relatively low - i.e., either 10 or 20 patients. Studies mention that as the cluster size increases, results from a random and fixed effects PS model should coincide - for example, see balance results from a fixed and random effects model in application in [54]. To estimate the PS weights we either: (i) integrated over the random effects distribution or (ii) predicted the EBE for the random effects and plugged them into the individual probabilities of treatment receipt. For the former, we obtained PS weights that could be interpreted as marginal on the clusters, hence balance on the cluster characteristics should be questionable. For the latter, EBEs are obtained from the assumed prior distribution of the random effects and the likelihood of the data conditional on the random effects [66]. Therefore, these estimates borrow information from across clusters. The influence of the assumed prior distribution on the final EBE value is called shrinkage. Shrinkage may be greater in cases where, e.g., the cluster itself cannot provide much information due to low cluster size, hence the final value will be 'shrunk' towards the cluster mean (i.e., the mean of the assumed prior distribution for the random effects, which across our settings was assumed to be zero). Shrinkage can be thought of as an advantage of the random effects models when applied to model the outcome in presence of small cluster sizes; however, when applying PS methods, it does not ensure balance on the cluster characteristics between the different individually assigned treatment levels when the cluster size is small.

The non-parametric clustered estimator eliminates confounding at the cluster level, regardless of the PS model applied; this is because, it calculates the ATE within cluster, where all cluster-level characteristics are constant; then, at a second step, it averages over clusters; however, for cases of too many small clusters, many of which, have only one treatment level being assigned, it cannot be

applied as it is. In such settings, either these clusters should be excluded (and the ATE should be calculated only among the overlapping clusters, which would mean loss of sample) or methods to group clusters of similar characteristics together should be used instead [50]. Nonetheless, in our simulation settings, we examined only cases for which, within-cluster positivity violations would be very unlikely to exist by design. We note that it has been demonstrated that for continuous outcomes and under the simplified setting of no covariates, the clustered estimator that applies a fixed effects PS model is equivalent with one that applies a marginal model; the marginal estimator that uses a fixed effects model for the PS is also shown to be equivalent to the previous two, under a balanced design. This makes all three estimators statistically consistent, and proves that, for continuous outcomes when one applies PS weighting estimators approximately unbiased estimation is achieved as long as cluster-level confounding is eliminated in at least one step of the PS analysis (i.e., either the PS model or the PS estimator itself) [54].

The doubly robust estimator applied in our simulations is the augmented inverse probability-of-treatment weighted estimator - $AIPTW$, initially proposed by Robins et. *al.* [71]. In our settings, we applied random effects models where the cluster-level confounder V, was omitted across analyses. Therefore, both the PS and the outcome models were misspecified. That could explain the bias introduced in the AIPTW estimates of the $p \quad ATE$. We note that, the applied AIPTW predicts the individual probabilities of the outcome (or of treatment receipt, for the PS model) that are marginal on the clusters, when integrating over the random effects distribution. In the case of plugging-in the EBEs, as in, e.g., eq. (3.12), technically, the individual probabilities within each cluster are calculated, before averaging over clusters to estimate the $p \quad ATE$.

Although bias was the focal point in our empirical evaluations of the compared estimators in the hierarchical setting, in chapter 5 we have provided two different methods to obtain approximate marginal variance formulas for propensity score weighting estimators when targeting the *ATE*, *ATT* and *ATO*. In the single level we have shown that not accounting for PS estimation results in either over or underestimation of the large-sample marginal variance estimates - depending on the targeted population. Variance of the examined estimators in the two-level setting could be estimated by the non-parametric bootstrap with resampling of the clusters; previous studies have followed this procedure, although it seems as more of an 'ad hoc' method, as no further justification of that particular choice could be found [54, 75]. Approximate variance formulas for such estimators (and whether possible to be derived), should be examined by specific case of estimator and evaluated empirically in contrast to bootstrap methods tailored for structured data sets, which is something not covered by this dissertation.

6.2 Strengths and limitations

Strengths

We summarize the key strengths of this work, into the following:

1. We solely focused on random effects models and marginal interpretation of the causal average treatment effect ($p \quad ATE$ herein), under omission of a cluster-level confounder, as these models are quite popular in the discipline of epidemiology - usually, when targeting conditional and descriptive effect measures, but not causal effects [67]. Additionally, issues with the exogeneity assumption have been rarely addressed in the literature of epidemiological applications [13]; therefore, epidemiologists tend to apply these without further justification or clear interpretation of the target estimands, just out of convenience and/or availability across statistical software.
2. The focus solely on random effects made comparisons between estimators that model the outcome *vs* the treatment more straightforward.
3. Although the PS estimators presented within this dissertation have been introduced and evaluated empirically before [54], empirical comparison between these and g-computation/standardization estimators combined with mixed effects was not performed within these studies; in addition, there is still a requirement for empirical studies illustrating the performance of such estimators under different settings; therefore, our investigations contributed in that aspect.
4. Importantly, we have provided reproducible R code both to generate and analyze the simulated data sets across our simulation settings, which enables reproducibility of results and can be found at sims-ISCB2023.
5. Finally, we have derived variance formulas for single-level IPTW estimators that are unifying to various target estimands and account for uncertainty in the PS estimation, providing valid approximate marginal variance estimates.

Limitations

We now summarize the main limitations of this work, into the following:

1. One weakness in our simulation settings across chapters 3 and 4, was that balance diagnostics were not examined within clusters for the simulated samples *before* and *after* weighting for the different PS weighted estimators.
2. Additionally, results for a number of DGMs[1] were not included in this thesis, due to time restrictions. Nonetheless, we assume inferences drawn should be consistent with those drawn from the DGMs included herein (see also nested loop plots in appendix C).

[1] 81 DGMs corresponding to codes 3,4,12,14,15,16,19,20,31,32,36,37,38,39,44,47,48,50,51,52,68,69,70,79,80,85,94, 96,109,113,119,121,124,125,134,139,141,148,158,164,182,186,193,197,201,202,205,206,209,214,217,218, 221,222,224,226,234,237,239,241,245,249,250,253,254,259,261,262,264,266,269,270,271,274,277,278,280, 281,282,285,286 were not included in *this version of the simulation results* presented in 4.4 and in Appendix C.

3. Moderate correlation was assumed between the random slopes and the random intercepts; for different values of correlation, results may vary. In the same manner, treatment effects did not vary between different clusters across DGMs - nonetheless, this was not in the scope of our work.
4. When predicting the individual probabilities of response (either when the response was the outcome or the treatment), we applied either the EBE or averaged over the prior distribution of the clusters; however, we did not include in the comparison, application of the estimates marginalized over the posterior distribution of the random effects, which would be of interest to test.
5. Due to time restrictions, variance performance was not examined. Our simulations illustrated performance under specific scenarios of data generating models; analytical calculations of bias under omission of a cluster level confounder could work as a complementary tool to investigate further performance of these estimators; however, it becomes quite intricate once the outcome ceases to be continuous and in presence of covariates [54].
6. Lastly, more research is required to clarify why EBE versions of the marginal IPTW and AIPTW performed better compared to their population averaged equivalents (and whether this result can be generalized to larger cluster sizes too) - although we have provided an intuition on the matter (see 6.1).

6.3 Matters of future research

During the course of my PhD I have identified potential avenues of future research:

This dissertation has focused on binary treatments assigned at the individual level. An interesting future avenue could be settings with continuous treatments, as these are relevant to cancer research, while methodological matters arise - such as the formulation of the stabilized weights, which are suggested in the single level setting for continuous covariates over the unstabilized ones, when applying PS weighting [75]. Treatments assigned at the cluster level could be of interest too (see 2.3.1). For cases where we need to model effects of continuous covariates that have complex polynomial (or smoother) effects, a class of models called generalized additive models (GAMs) is gaining popularity [95]. Lastly, Bayesian approaches have also been introduced for hierarchical data (e.g., see [29]), although, our focus was solely on frequentist methods.
In our simulation studies, we applied the *AIPTW* that is a DR estimator. To investigate other DR estimators, such as *TMLE* [11, 58], and their comparative performance in the two-level setting could be an interesting future avenue. We note that recent progress for treatments assigned at the cluster level has been made [11]. *TMLE* is reported to have desired properties, such as robustness towards model misspecification compared to more 'parametric' methods.

Different estimators were applied in the two-level setting to target the $p \quad ATE$; it would be interesting to perform similar investigations when targeting other populations, such as the *ATT*.
In our empirical experiments, we examined g-computation methods when mixed effects models are applied to model the outcome under scenarios of individual-level covariates being correlated with the omitted cluster-level covariate, which is known to be introducing bias in the cluster-specific regression coefficient estimates. Recent modeling techniques have been proposed when modeling the outcome to relax this assumption [13]. Interesting would be to explore under simulation studies whether combining such estimators with the standardization step of g-computation would decrease bias in the $p \quad ATE$ even more. These approaches would be useful in settings where the exogeneity assumption of the random effects is violated and at the same time, effects of cluster-level covariates are of interest to estimate from the (outcome) model.
Another avenue of future research would be to extend the examined estimators to models for time-to-event data and time-dependent covariates, as different challenges arise for the causal inference framework [42]; for instance, for Cox models, recent results show, that just combining treatment and outcome models into one, does not necessarily provide double robustness - e.g., see [27]. Link functions for the potential outcomes model, such as the c-loglog could be of interest too.
Our main finding from the short Monte Carlo simulations we conducted under the null in chapter 3, was that the use of random effects to model the outcome under the omission of a cluster-level confounder (although not eliminated completely), decreased bias in the estimation of the $p \quad ATE$. Although further research is needed to confirm this is valid across different simulation settings, this result was consistent across simulation settings where we assumed a moderate treatment effect in chapter 4. Possible steps following that investigation would be to test whether we obtain similar conclusions when the treatment effect varies between different clusters (i.e., treatment effect heterogeneity). We note that the tested methods were empirically examined only in terms of bias; a critical next step is to explore variance performance of these estimators. This should be applied both for the simulation settings examined in Sections 3.3 and 4.4. Matters of future interest related to our unifying large sample IPTW variance formula in the single- level setting were mentioned in 5.3.
Finally, a crucial next step should be to apply all estimators presented in Chapters 3 and 4 to real-world cancer data sets, including (but not limited to) studies that aim to unfold effects on observed cancer inequalities.

6.4 Conclusion

Practical implications for epidemiologists

Random effects can be seen as a compromise between fixed effects and marginal models under unobserved cluster-level confounding. Under our simulation settings (with cluster sizes no more than 20 individuals on average per cluster), these performed better when combined with g-computation than when with PS weighted estimators - note that although the inclusion of random effects within our

simulations seemed to reduce estimated bias, random effects models cannot provide fully unbiased estimates when important cluster-level confounders are omitted from analysis due to violation of the exogeneity assumption. In terms of interpretation, we advise caution, as causal interpretations of regression coefficients in a multilevel model can be misleading, as it has already been pointed out in the literature (e.g., see [66, 28]). For PS methods, it would be advisable to eliminate cluster-level confounding non-parametrically, when possible (see non-parametric clustered IPTW); otherwise, one should apply alternative methods to achieve cluster-level balance such as fixed effects, if applicable - however, it has been suggested that these models for the PS may cause variance inflation [54]. These results are in line with Li et *al.* [54] that suggest cluster-level confounding must be accounted for in at least one step of the PS analysis (either by the PS model or the PS estimator itself).

Overall contribution of this work

This dissertation provided a brief review of causal inference methods for observational data; it evaluated via Monte Carlo experiments the performance of g-computation, *IPTW* and *AIPTW* in terms of bias when combined with random effects models under omission of a cluster-level confounder; it further examined estimands and interpretation of the resulting estimates; finally, it offered unified analytical formulas for the derivation of the marginal large-sample IPTW variance via two estimating techniques that are widely applied in practice, i.e., linearization and M-estimation, while targeting different estimands in the single-level setting, i.e., the ATE, ATT and ATO. Any related code to reproduce all the simulations performed was made publicly available and can be found at the respective repositories on GitHub at sims-gcomp and sims-ISCB2023. This contributes to research dissemination, which falls under the general scope of good research practice. Our overarching conclusion is that further systematic research is required to unfold and illustrate the behaviour of the methods examined herein, under different settings that are plausible to be encountered in epidemiological practice.

References

[1] Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35.

[2] Abadie, A. and Imbens, G. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557.

[3] Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley, second edition.

[4] Angrist, J., Imbens, G., and Rubin, D. (1993). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

[5] Arpino, B. and Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, 55(4):1770–1780.

[6] Austin, P. and Lee, D. (2020). Estimating the net benefit of improvements in hospital performance: G-computation with hierarchical regression models. *Medical Care*, 58(7):651–657.

[7] Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics*, 10(2):150–161.

[8] Austin, P. C. (2016). Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine*, 35:5642–5655.

[9] Austin, P. C. and Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in Medicine*, 33:4306–4319.

[10] Austin, P. C. and Urbach, D. R. (2013). Using g-computation to estimate the effect of regionalization of surgical services on the absolute reduction in the occurrence of adverse patient outcomes. *Medical Care*, 51:797–805.

[11] Balzer, L. B., Zheng, W., Van der Laan, M. J., and Petersen, M. L. (2019). A new approach to hierarchical data analysis: Targeted maximum likelihood estimation for the causal effect of a cluster-level exposure. *Statistical Methods in Medical Research*, 28(6):1761–1780.

[12] Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

[13] Bates, M. D., Castellano, K. E., Rabe-Hesketh, S., and Skrondal, A. (2014). Handling correlations between covariates and random slopes in multilevel models. *Journal of Educational and Behavioral Statistics*, 39:524–549.

[14] Belot, A., Fowler, H., Njagi, E. N., Luque-Fernandez, M.-A., Maringe, C., Magadi, W., Exarchakou, A., Quaresma, M., Turculet, A., Peake, M. D., Navani, N., and Rachet, B. (2019). Association between age, deprivation and specific comorbid conditions and the receipt of major surgery in patients with non-small cell lung cancer in england: A population-based study. *Thorax*, 74(1):51–59.

[15] Boos, D. D. and Stefanski, L. A. (2013). *Essential Statistical Inference: Theory and Methods*. Springer Texts in Statistics, Springer New York.

[16] Cafri, G., Wang, W., Chan, P. H., and Austin, P. C. (2018). A review and empirical comparison of causal inference methods for clustered observational data with application to the evaluation of the effectiveness of medical devices. *Statistical Methods in Medical Research*, 28:3142–3162.

[17] Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in Medicine*, 19:1141–1164.

[18] Chang, T.-H. and Stuart, E. A. (2022). Propensity score methods for observational studies with clustered data: A review. *Statistics in Medicine*, 41(18):3612–3626.

[19] Charvat, H., Remontet, L., Bossard, N., Roche, L., Dejardin, O., Rachet, B., Launoy, G., and Belot, A. (2016). A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Statistics in Medicine*, 35(18):3066–3084.

[20] Cole, S. R. and E., F. C. (2009). The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, 20(1):3–5.

[21] Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664.

[22] Daniel, R., Zhang, J., and Farewell, D. (2021). Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal*, 63:528–557.

[23] Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424.

[24] Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25(2):193–203.

[25] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/ CRC.

[26] Fan, L., Kari, L. M., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400.

[27] Gabriel, E. E., Sachs, M. C., Waernbaum, I., Goetghebeur, E., Blanche, P. F., Vansteelandt, S., Sjölander, A., and Scheike, T. (2024). Propensity weighting plus adjustment in proportional hazards model is not doubly robust. *Biometrics*, 84(3).

[28] Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, 48(3):432–435.

[29] Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

[30] Goetghebeur, E., le Cessie, S., Stavola, B. D., Moodie, E. E., and and, I. W. (2020). Formulating causal questions and principled statistical answers. *Statistics in Medicine*, 39:4922–4948.

[31] Greenland, S., Pearl, J., and Robins, J. M. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46.

[32] Greenland, S. and Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15(3):413–419.

[33] Greenland, S. J. P. and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology (Cambridge, Mass.)*, 10(1):37–48.

[34] Greifer, N. and Stuart, E. A. (2021). Choosing the causal estimand for propensity score analysis of observational studies. *arXiv preprint arXiv:2106.10577*.

[35] Griswold, M. E., Swihart, B. J., Caffo, B. S., and Zeger, S. L. (2013). Practical marginalized multilevel models. *Stat*, 2:129–142.

[36] Hajage, D., Chauvet, G., Belin, L., Lafourcade, A., Tubach, F., and Rycke, Y. D. (2018). Closed-form variance estimator for weighted propensity score estimators with survival outcome. *Biometrical Journal*, 60(6):1151–1163.

[37] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69:383.

[38] Hayes, R. J. and Moulton, L. H. (2009). *Cluster Randomised Trials*. Chapman and Hall.

[39] Heagerty, P. J. and Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statistical Science*, 15(1):1–19.

[40] Hedeker, D., du Toit, S. H. C., Demirtas, H., and Gibbons, R. D. (2018). A note on marginalization of regression parameters from mixed models of binary outcomes. *Biometrics*, 74(1):354–361.

[41] Hernán, M. A. (2004). A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health*, 58(4):265–271.

[42] Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology*, 21(1):13–15.

[43] Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What if*. Chapman and Hall.

[44] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

[45] Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86:4–29.

[46] Jonsson Funk, M., Westreich, D., Wiesen, C., Sturmer, T., Brookhart, M. A., and Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767.

[47] Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22(4):523 – 539.

[48] Kim, G.-S., Paik, M. C., and Kim, H. (2017). Causal inference with observational data under cluster-specific non-ignorable assignment mechanism. *Computational Statistics & Data Analysis*, 113:88–99.

[49] Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29:337–346.

[50] Lee, Y., Nguyen, T. Q., and Stuart, E. A. (2021). Partially pooled propensity score models for average treatment effect estimation with multilevel data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184:1578–1598.

[51] Leon, A. C. and Hedeker, D. (2007). Quintile stratification based on a misspecified propensity score in longitudinal treatment effectiveness analyses of ordinal doses. *Computational Statistics & Data Analysis*, 51:6114–6122.

[52] Li, F. and Thomas, L. E. (2018). Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*, 188(1):250–257.

[53] Li, F., Zaslavsky, A. M., and Landrum, M. B. (2007). Propensity score analysis with hierarchical data. *In: Proceedings of the American Statistical Association Joint Statistical Meetings; July 29 - August 2, 2007; Salt Lake City, UT*.

[54] Li, F., Zaslavsky, A. M., and Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, 32(19):3373–3387.

[55] Li, L. and Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. *The International Journal of Biostatistics*, 9(2):215–234.

[56] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

[57] Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960.

[58] Luque-Fernandez, M. A., Schomaker, M., Rachet, B., and Schnitzer, M. E. (2018). Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine*, 37(16):2530–46.

[59] McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403–425.

[60] Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102.

[61] Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32.

[62] Pavlou, M., Ambler, G., Seaman, S., and Omar, R. Z. (2015). A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes. *BMC Medical Research Methodology*, 15(1):59–59.

[63] Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82:702–710.

[64] Pearl, J. (2010). On the consistency rule in causal inference: Axiom, definition, assumption, or theorem? *Epidemiology*, 21:872–875.

[65] Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., and Van der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res*, 21(1):31–54.

[66] Rabe-Hesketh, S. and Skrondal, A. (2012a). *Multilevel and Longitudinal Modeling Using Stata Volume I: Continuous Responses*. Stata Press Publication, third edition.

[67] Rabe-Hesketh, S. and Skrondal, A. (2012b). *Multilevel and Longitudinal Modeling Using Stata Volume II: Categorical Responses, Counts, and Survival*. Stata Press Publication, third edition.

[68] Rizopoulos, D. (2023). *GLMMadaptive: Generalized Linear Mixed Models using Adaptive Gaussian Quadrature*. https://github.com/drizopoulos/GLMMadaptive.

[69] Robbins, M. W., Griffin, B. A., Shih, R. A., and Slaughter, M. E. (2019). Robust estimation of the causal effect of time-varying neighborhood factors on health outcomes. *Statistics in Medicine*, 39:544–561.

[70] Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512.

[71] Robins, J., Rotnitzky, A., and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

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

[73] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[74] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 56:688–701.

[75] Schuler, M., Chu, W., and Coffman, D. (2016). Propensity score weighting for a continuous exposure with multilevel data. *Health Services and Outcomes Research Methodology*, 16(4):271–292.

[76] Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., and Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety*, 17:546–555.

[77] Sherman, M. and le Cessie, S. (1997). A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics - Simulation and Computation*, 26:901–925.

[78] Skrondal, A. and Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 172(3):659–687.

[79] Smith, M., Maringe, C., Rachet, B., Mansournia, M., Zivich, P., Cole, S., and Luque-Fernandez, M. (2020). Tutorial: Introduction to computational causal inference using reproducible Stata, R and Python code. *Statistics in Medicine*, 41(2):407–432.

[80] Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. *Statistical Science*, 5(4):465–472.

[81] Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician*, 56(1):29–38.

[82] Stewart, I., Aamir Khakwani, R. B. H., Beckett, P., Borthwick, D., Tod, A., Leary, A., and Tata, L. J. (2018). Are working practices of lung cancer nurse specialists associated with variation in peoples' receipt of anticancer therapy? *Lung Cancer*, 123:160–165.

[83] Stroud, A. H. and Sechrest, D. (1966). *Gaussian Quadrature Formulas*. Englewood Cliffs, NJ: Prentice Hall.

[84] Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

[85] Tsiatis, A. A. and Davidian, M. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):569–573.

[86] Tuerlinckx, F., Rijmen, F., Verbeke, G., and Boeck, P. D. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, 59:225–255.

[87] Van der Laan, M. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.

[88] Van der Laan, M. J. and S., R. (2011). *Targeted learning: Causal inference for observational and experimental data*. New York: Springer Series in Statistics.

[89] VanderWeele, T. J. (2008). Ignorability and stability assumptions in neighborhood effects research. *Statistics in Medicine*, 27(11):1934–1943.

[90] VanderWeele, T. J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology*, 20(6):880–883.

[91] VanderWeele, T. J. (2010). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological Methods & Research*, 38:515–544.

[92] Westreich, D. and Cole, S. R. (2010). Invited commentary: Positivity in practice. *American Journal of Epidemiology*, 171(6):674–677.

[93] Williamson, E. (2007). *Inference from estimators of exposure effects obtained by stratification on the propensity score*. PhD thesis, London School of Hygiene and Tropical Medicine.

[94] Williamson, E. J., Forbes, A., and White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine*, 33(5):721–737.

[95] Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Taylor & Francis Inc, 2nd edition.

[96] Yang, S. (2018). Propensity score weighting for causal inference with clustered data. *Journal of Causal Inference*, 6(2):688–701.

[97] Yuan, Y. and Little, R. J. A. (2007). Model-based estimates of the finite population mean for two-stage cluster samples with unit non-response. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 56:79–97.

[98] Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44(4):1049.

[99] Zhao, Z. (2005). Sensitivity of propensity score methods to the specifications. *Economics letters*, 98(3):309–319.

[100] Zhou, Y., Matsouaka, R. A., and Thomas, L. (2020). Propensity score weighting under limited overlap and model misspecification. *Statistical Methods in Medical Research*, 29(12):3721–3756.

Appendix A

Appendix of Research Paper I

APPENDIX OF 'A NOTE ON THE ESTIMATION OF MARGINAL CAUSAL EFFECTS WITH MIXED EFFECTS MODELS'

Andriana Kostouraki, Clémence Leyrat, Aurélien Belot

A.1. Description of the simulation study

A.1.1 Aim

The aim of our simulation study ([1]) was to illustrate the performance of mixed e□ects models under di□erent data generating scenarios of the "unmeasured context" - i.e., when important cluster-level covariates are omitted from the analysis model ([2]).

A.1.2 Data generating mechanisms and estimands

We simulated □ = 5,000 patients for a number of simulation replicates equal to □□□□ = 1,000. We randomly generated two individual-level confounders □1 □ □(0, 1) and □2 □ □□□□□□□□□(0.5) and a cluster-level covariate, □ □ □(0, 1). When □1 was correlated with □, we set: □1 □= □ □ □ + $\sqrt{1 - □^2}$ □ □, with □ □ □(0, 1) and correlation □ □ {0.3, 0.5, 0.8}. We assumed a balanced design of 100 clusters of 50 patients. All scenarios were examined under the null hypothesis. We considered the data generating mechanisms (DGMs) below:

- DGM 1: □1, □2, individual-level and □ cluster-level confounders; no correlation among covariates; random intercept treatment and outcome models.

  Treatment model:

  $$□□□□□□(\{□□(□□□ = 1|□□0, □1□□, □2□□, □□\}) = (□0 + □□0) + □1□1□□ + □2□2□□ + □3□□ \quad (1)$$

  where $(□0, □1, □2, □3) = (1, 0.2, -0.5, 0.6)$ and □□0 □ □(0, 1).
  Outcome model: $□□□□□\{□□(□□□ = 1|□□0, □□□, □1□□, □2□□, □□)\} = (□0 + □□0) + □□□□□ + □1□1□□ + □2□2□□ + □3□□$, with $(□, □1, □2, □3) = (0, 0.2, -1, 2)$ and □□0 □ □(0, 1). For □0 equal to −3.2, −1, 0.1 we obtained marginal event rates of 10%, 30% and 45%, respectively (this was empirically computed via trial and error of di□erent values for □0 in random draws of 10^7 observations in total). Marginal treatment prevalence was around 65% across DGMs.

- DGM 2: as in DGM 1, but □ correlated with □1.

We targeted both the *marginal or population* average treatment e□ect, i.e., p-ATE and the cluster-specific (but, marginal on the covariates) average treatment e□ect, i.e., cs-ATE. We are interested in how (□) the p-ATE and the cs-ATE estimates are impacted under omission of a cluster-level confounder, (□□) the two approaches [1] to estimate the p-ATE comparatively perform and (□□□) ignoring clustering altogether a□ects the p-ATE estimates.

A.1.3 Methods and performance measures

To analyse the simulated data sets, we considered an: (□) unadjusted generalized linear model (unadjusted - GLM), (□□) unadjusted generalized linear mixed e□ect model with a random intercept (unadjusted - GLMM), (□□□) adjusted generalized linear model for all covariates (fully adjusted - GLM), (□□) adjusted generalized linear mixed e□ect model for all covariates with a random intercept (fully adjusted - GLMM), (□) adjusted generalized linear model for individual-level covariates (adjusted - individual GLM) and, (□□) adjusted generalized linear mixed e□ect model for individual-level covariates with a random intercept (adjusted - individual GLMM).

When it comes to performance measures, we primarily focused on bias, as the aim of this study was to showcase to what extent random e□ects may (or may not) be used as proxy to prevent bias when a cluster-level covariate is not included in the analysis model; we showcase additional results on coverage, for 100 simulations and □ = 100 bootstrap replicates for each simulated data set (see at the end of this Appendix).

[1] these two appproaches are: (i) plugging-in the EBEs and (ii) integrating over the random e□ects distribution.

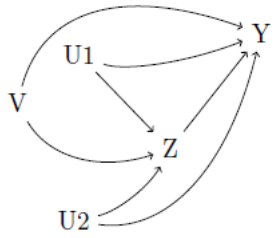


FIGURE 1 Directed acyclic graph (DAG): □ is assumed to predict both the treatment and the outcome (confounder). Minimal su□cient adjustment sets for estimating the e□ect of □ on □ □ □$_1$, □$_2$, □ .

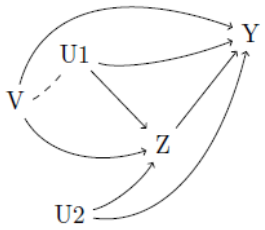


FIGURE 2 Graphical depiction of correlation between □ and □$_1$ in dotted lines. Minimal su□cient adjustment sets for estimating the e□ect of □ on □ □ □$_1$, □$_2$, □ .

Therefore, we aimed to explore if a GLMM analysis, adjusted *only* for the individual-level covariates, allows estimating the p-ATE as successfully as a fully adjusted GLMM when □ is omitted, under the null. We also investigated what may be the consequences of ignoring a hierarchical structure. Reproducible code to simulate and analyse the data can be found in the corresponding appendix. All simulations were performed in R, version 4.3.0.

Lastly, to ensure we obtain a reasonable distribution for Y, we may draw a relatively large sample, e.g.: 10, 000, 000 observations in total, for a single simulation, and check the marginal event rate of □ , as well as the parameter estimates we obtain from an analysis model that corresponds to the true data generating model approximately coincide. For example, for values of □ = 50, □ = 200 □ 10^3, □$_0$ = 0.1, □ = 0.5, seed= 20230523, for a single simulation, we get an event rate of □ 45%, etc.

A.1.4. Additional simulation results

A.1.4.1 Bias

We present additional results when the correlation between □ and □$_1$ is assumed 0.3 and 0.8, for marginal event rates 45% (see Tables 1 , 2). Conclusions on bias results for marginal event rates of 10% and 30% were similar to those for 45% and are suppressed for brevity.

A.1.4.2 Coverage

Below, for ease in visualization, we present results on coverage rate estimates for 100 simulations and 100 bootstrap replicates. For the standard error estimates within each simulated sample, we applied the non-parametric bootstrap across analyses methods, by sampling the clusters with replacement. We constructed the 95% normal-based CIs. The number of the seed was 20230523. The results below correspond to □$_0$ = 0.1, which translates to a marginal event rate that is approximately 45% and an assumed correlation between □ and □$_1$ equal to 0.5. Results for the rest combinations of parameter values were similar and suppressed for brevity.

TABLE 1 100×Bias for: DGM 2 (confounder of and , random intercept), = 100 clusters of = 50 individuals, $\beta_3 = 2$, $\beta = 0$, $\sigma = 0.3$, correlation of and 1 and = 1,000 replications.

	ATE-cs	ATE-ebe	ATE-integ
unadjusted - GLM	-	15.90	15.90
unadjusted - GLMM	4.42	2.71	2.66
fully adjusted - GLM	-	-0.06	-0.06
fully adjusted - GLMM	-0.01	-0.01	-0.01
adjusted - individual GLM	-	11.78	11.78
adjusted - individual GLMM	0.75	0.47	0.47

TABLE 2 100×Bias for: DGM 2 (confounder of and , random intercept), = 100 clusters of = 50 individuals, $\beta_3 = 2$, $\beta = 0$, $\sigma = 0.8$, correlation of and β_1 and = 1,000 replications.

	ATE-cs	ATE-ebe	ATE-integ
unadjusted - GLM	-	17.99	17.99
unadjusted - GLMM	4.13	2.47	2.43
fully adjusted - GLM	-	-0.04	-0.04
fully adjusted - GLMM	0.02	0.02	0.01
adjusted - individual GLM	-	4.66	4.66
adjusted - individual GLMM	0.82	0.51	0.53

Briefly, within the unadjusted analyses, the GLM had the worst performance in terms of coverage, with the random intercept unadjusted analyses having a better performance both for the cs- and the p-ATE, but not above 42% (Figures 3 a, 3 b, 3 c, 3 d). All the fully adjusted analyses achieved a coverage rate that was closest to the advertised as anticipated (Figures 4 a, 4 b, 4 c, 4 d). With an increase in the number of the bootstrap replicates, we expect to have coverage rates almost equal to 95%. Lastly, within the analyses adjusted only for the individual-level covariates: the GLM analysis performed the worst (8% coverage), whereas all the random intercept analyses had coverage values much closer to - although still lower than - the nominal coverage rate (5 a, 5 b, 5 c, 5 d).

A.2. Identifiability assumptions for hierarchical structures and treatment assigned at the individual level

Herein, we focus on two-level structures. Let us assume an observational cohort study. Consider a sample or population of n patients that consist of H clusters, with the cluster indicator = 1, 2, ..., H and the cluster specific size . The total sample size will be $n = \sum_{i} n_i$. We denote U_{ij} a vector of unit level covariates for patient within cluster and V_i a vector of cluster level covariates for the same cluster.

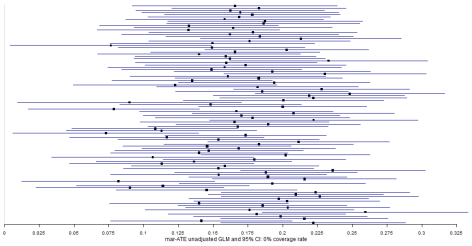
We adjust the identifiability assumptions for single-level settings to the following:

1. No interference, i.e., the potential individual outcome $Y_{ij}(z)$ for the j^{th} patient nested within the i^{th} cluster is independent of the treatment assigned to any other individual nested within a cluster , with ≠ , and = or ≠ (i.e., "the outcomes for each unit are una ected by the treatment assignments of other units whether within or across clusters", [3]).

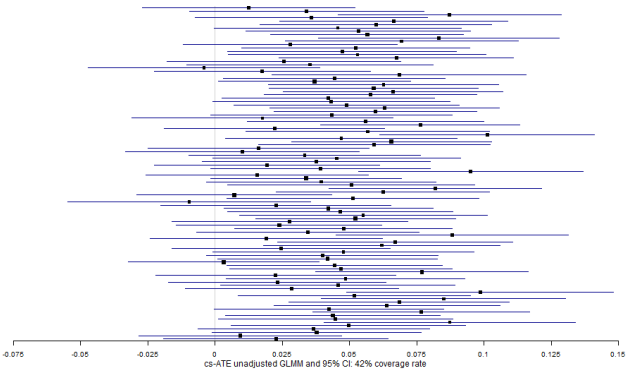
2. Consistency, i.e., for the patients that actually received treatment of level their observed outcome is the same as what it would have been had they received treatment level via the hypothetical intervention we have in mind (or in other words, there are no multiple versions of treatment and observing is the same as imposing):

$$Z_{ij} = z \;\Rightarrow\; Y_{ij} = Y_{ij}(z) \tag{2}$$

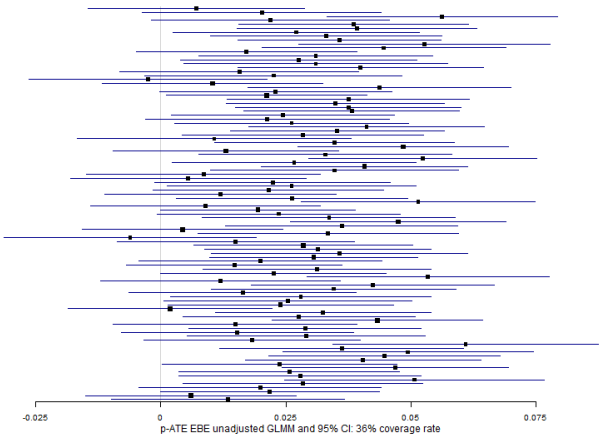
3. Conditional exchangeability (or ignorability or no unmeasured confounding), i.e., the potential outcomes (z) are conditionally independent of the treatment Z given the pre-treatment measured covariates X for every treatment level



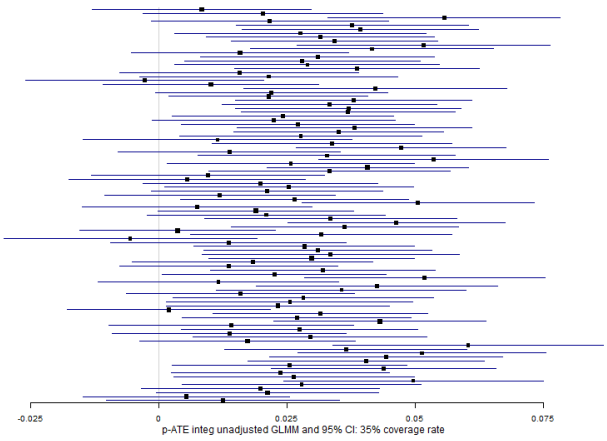
(a)



(b)

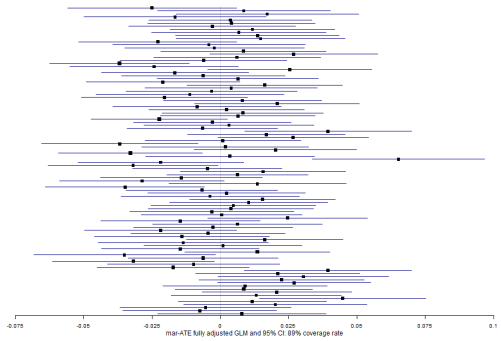


(c)

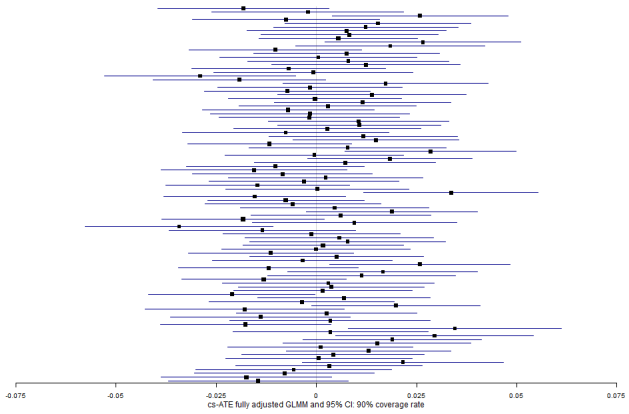


(d)

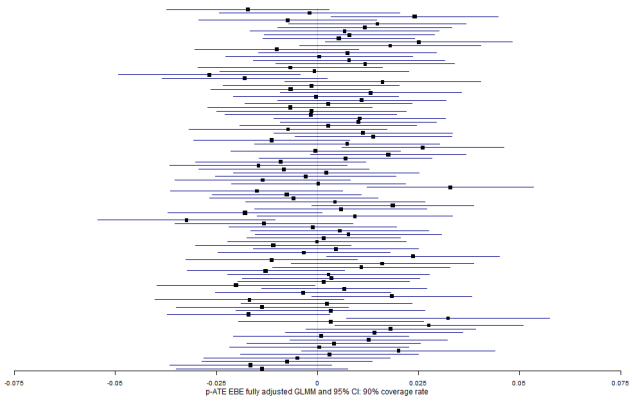
FIGURE 3 Coverage rate: unadjusted analyses; $\square_{\square\square\square}$ = 100, $\square$ = 100 bootstrap replicates, non-parametric bootstrap, $\square_0$ = 0.1, $\square$ = 0.5.



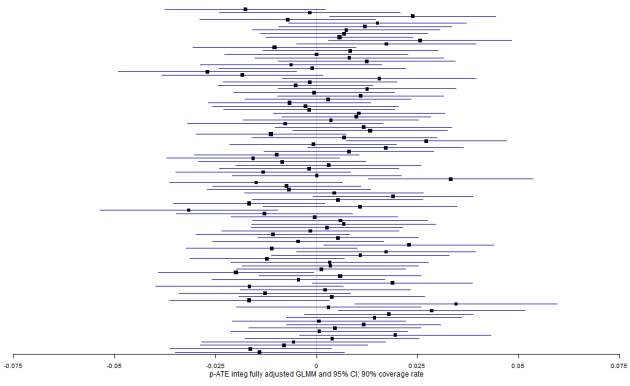
(a)



(b)

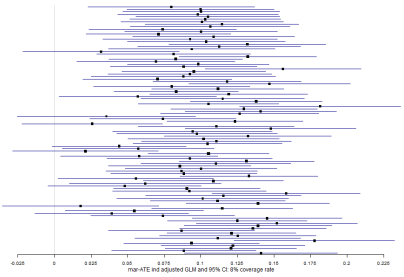


(c)

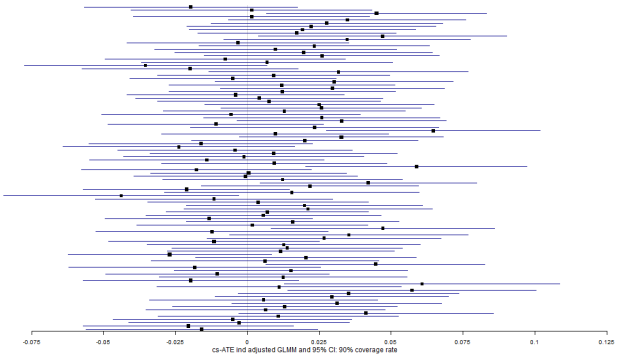


(d)

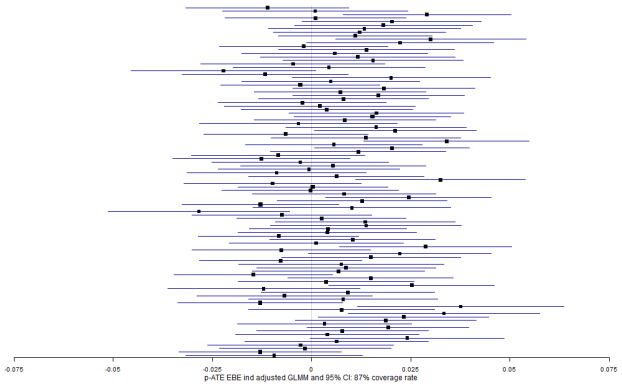
FIGURE 4 Coverage rate: fully adjusted analyses; $\square_{\square\square\square}$ = 100, $\square$ = 100 bootstrap replicates, non-parametric bootstrap, $\square_0$ = 0.1, $\square$ = 0.5.



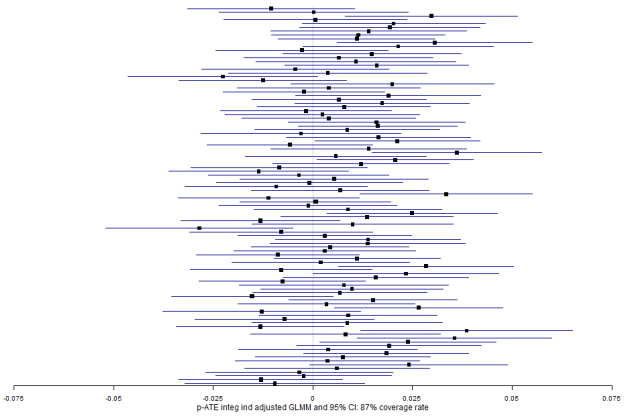
(a)



(b)



(c)



(d)

FIGURE 5 Coverage rate: analyses adjusted for individual covariates; $\square_{\square\square\square}$ = 100, $\square$ = 100 bootstrap replicates, non-parametric bootstrap, $\square_0$ = 0.1, $\square$ = 0.5.

z:

$$\square_{\square\square}(\square) \text{ m } \square_{\square\square}|X_{\square\square}, \square\square \quad (3)$$

Where the symbol m indicates statistical independence and the vector $X_{\square\square} = (U_{\square\square}, V_{\square})$ this time contains vectors $U_{\square\square}$ and $V_{\square\square}$ of unit-level and cluster-level covariates respectively. So, again, this is equivalent to saying that treatment allocation is *at random within di erent combinations of observed individual- and cluster-level characteristics*.

4. Positivity: i.e., every patient has non zero probability of receiving treatment of level z, for every di□erent level z of treatment Z:

$$\square\square(X_{\square\square} = x) > 0 \text{ } \square \text{ } 0 < \square\square(\square_{\square\square} = \square|X_{\square\square} = x) < 1, \square\square, x \quad (4)$$

where $X_{\square\square} = (U_{\square\square}, V_{\square})$ is the vector of individual- and cluster-level covariates.

A.3. Integral for the marginal or population average probabilities of response (for random intercepts and slopes models)

([4],[5]):

$$\hat{\square}^{\square\square}_{\square\square}(\square) = \int_{\square\square\square} \square\square\square\square(\square'_{\square\square}\hat{\square}^{\square\square} + \square'_{\square\square}\square_{\square})\square\square(\square), \square = 0, 1 \quad (5)$$

where $\square\square\square\square(\square) = \frac{\square\square\square(\square)}{1+\square\square\square(\square)}$, $\hat{\square}^{\square\square}$ the cluster-specific coe□cient estimates from a logistic mixed-e□ects model and $\square_{\square}$ are the random e□ects, with $\square_{\square} \sim \square(\square, \square_{\square})$ - where $\square_{\square}$ the variance-covariance matrix of the random e□ects. Additionally, $\square$ is the support space of the distribution of the random e□ects, $\square_{\square\square}$ the vector of covariate values (including the treatment $\square$), and, $\square_{\square\square}$, the vector of covariate values for the random e□ects. $\square$, is the cumulative probability density function of the random e□ects.

References

1. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 2019; 38(11): 2074–2102. doi: 10.1002/sim.8086

2. Arpino B, Mealli F. The specification of the propensity score in multilevel observational studies. *Computational Statistics and Data Analysis* 2011; 55(4): 1770–1780. doi: 10.1016/j.csda.2010.11.008

3. Li F, Zaslavsky AM, Landrum MB. Propensity score weighting with multilevel data. *Statistics in Medicine* 2013; 32(19): 3373–3387. doi: 10.1002/sim.5786

4. Skrondal A, Rabe-Hesketh S. Prediction in Multilevel Generalized Linear Models. *Journal of the Royal Statistical Society Series A: Statistics in Society* 2009; 172(3): 659–687. doi: 10.1111/j.1467-985x.2009.00587.x

5. Griswold ME, Zeger SL. On Marginalized Multilevel Models and their Computation. 2004; Johns Hopkins University, Dept. of Biostatistics Working Papers, Working Paper 99.

□

Appendix B

Appendix of Research Paper II

On variance estimation of the inverse probability-of-treatment weighting estimator: a tutorial for different types of propensity score weights
Appendix I

General formula for different types of propensity score weights

Andriana Kostouraki, David Hajage, Bernard Rachet, Elizabeth J. Williamson,

Guillaume Chauvet, Aurélien Belot, Clémence Leyrat

February 12, 2024

Contents

1 General formula for different types of propensity score weights

Herein we showcase the derivation of the IPTW variance estimator via M-estimation and its equivalence to the one obtained via linearization. As logistic regression to model the PS is the most frequently applied one, in this Appendix we focus on this specific case.

1.1 Notation

We use the same notation as in the main article. Normally, vectors are denoted in **bold**:

- $Z_i \quad \{0, 1\}$: observed treatment value for individual i
- Y_i: observed outcome value for individual i
- n: total sample size
- p: total number of covariates
- $\mathbf{x}_i = (1, X_{1i}, ..., X_{pi})^T$ the $(p + 1)$-column vector of intercept and covariates for individual i, $i = 1, ..., n$
- $\mathbf{X}$: $n \times (p + 1)$ design matrix of the measured baseline characteristics (i.e. potential confounders):

$$\mathbf{X} = \begin{bmatrix} 1 & X_{(1)1} & \cdots & X_{(p)1} \\ 1 & X_{(1)2} & \cdots & X_{(p)2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{(1)i} & \cdots & X_{(p)i} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{(1)n} & \cdots & X_{(p)n} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_i \\ \vdots \\ \mathbf{x}_n \end{bmatrix}$$

  where $X_{(1)i}$ for instance represents the first covariate for the i^{th} participant and $\mathbf{x}_i = (1, X_{(1)i}, ..., X_{(p)i})$ represents the vector of observed covariates for the i^{th} participant. The first element of the vector is set to be equal to 1 to illustrate the intercept term that is applied in the regression models fitted in the subsequent segments.

- The estimated propensity score for individual i, with $\alpha = (\alpha_0, \alpha_1, ..., \alpha_p)^T$ the vector of propensity score parameters, is:

$$\hat{e_i} = g^{-1}(\mathbf{x}_i \, \hat{\alpha}) \tag{1.1}$$

  And, for the logistic regression, we have:

$$\hat{e}_i = \frac{\exp(\mathbf{x}_i^T \hat{\alpha})}{1 + \exp(\mathbf{x}_i^T \hat{\alpha})} \tag{1.2}$$

  with $\hat{\alpha}$, the maximum likelihood (ML) estimates.

- $\hat{w}_{\cdot i} = f_{\cdot}(\mathbf{x}_i^T \hat{\alpha})$ the estimated weights, $\cdot = 0$ if $Z = 0$, 1 otherwise.
- The estimated mean outcomes in the intervention and control groups respectively, are simply the weighted averages:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} \hat{w}_{1i} Z_i Y_i}{\sum_{i=1}^{n} \hat{w}_{1i} Z_i} \tag{1.3}$$

and

$$\hat{\mu}_0 = \frac{\sum_{i=1}^{n} \hat{w}_{0i} (1 - Z_i) Y_i}{\sum_{i=1}^{n} \hat{w}_{0i} (1 - Z_i)} \tag{1.4}$$

1.2 Weights

- ATE: $\hat{w}_{1i}^{ATE} = \frac{1}{\hat{e}_i}$ if $Z = 1$ and $\hat{w}_{0i}^{ATE} = \frac{1}{1-\hat{e}_i}$ if $Z = 0$
- ATT: $\hat{w}_{1i}^{ATT} = 1$ if $Z = 1$ and $\hat{w}_{0i}^{ATT} = \frac{\hat{e}_i}{1-\hat{e}_i}$ if $Z = 0$
- Overlap weights: $\hat{w}_{1i}^{OW} = 1 - \hat{e}_i$ if $Z = 1$ and $\hat{w}_{0i}^{OW} = \hat{e}_i$ if $Z = 0$
- Matching weights: $\hat{w}_{1i}^{MW} = \frac{min(\hat{e}_i, 1-\hat{e}_i)}{\hat{e}_i}$ if $Z = 1$ and $\hat{w}_{0i}^{MW} = \frac{min(\hat{e}_i, 1-\hat{e}_i)}{1-\hat{e}_i}$ if $Z = 0$

1.3 Causal treatment effect measures

- $\delta_1 = \mu_1 - \mu_0$ is the mean or risk difference
- $\delta_2 = \log \frac{\mu_1}{\mu_0}$ is the logarithm of the marginal risk ratio
- $\delta_3 = \log \frac{\mu_1/(1-\mu_1)}{\mu_0/(1-\mu_0)}$ is the logarithm of the marginal odds ratio

1.4 M-estimation

The aim is to extend the proposed by Williamson *et al.* ATE variance estimator to several types of weights. In this subsection to derive the final variance estimator, we use M-estimation and simple matrix multiplication. Note here that for ATE weights, some simplifications can be applied; however, this is not the case for the rest types of weights (and their derivatives) in the formulae in the main text.

Our ultimate goal is to estimate $Var(\hat{\delta}_j)$, $j = 1, 2, 3$. The idea is that δ_j is a function of μ_1 and μ_0; hence, we need to stack the estimating equations that are equivalent to estimating $\theta_0 = (\mu_1, \mu_0, \alpha^T)^T$. Obviously, one way is to replace the estimated individual PS values to calculate the IPTW estimators $\hat{\mu}_1$ and $\hat{\mu}_0$ (eq 3 and 4, section 2.2). An alternative to the latter would be to solve the corresponding estimating equations that can be derived based on the so-called M-estimation theory (Stefanski and Boos, 2002). Therefore, we need to solve:

$$\sum_{i=1}^{n} u(\theta; Y_i, Z_i, x_i) = 0 \tag{1.5}$$

with $u(\theta; Y_i, Z_i, x_i) = \begin{bmatrix} (Y_i - \mu_1)Z_i w_{1i} \\ (Y_i - \mu_0)(1 - Z_i)w_{0i} \\ x_i(Z_i - e_i) \end{bmatrix}$. The first two terms are scalars, the last is a $(p + 1)$-column vector; $\hat{\theta}$ is asymptotically normally distributed, with large-sample variance equal to: $n^{-1}A^{-1}BA^{-T}$ with $A = -\mathrm{E}\left[\frac{\partial u}{\partial \theta^T}\right]$ and $B = \mathrm{E}\left[uu^T\right]$. Therefore, we need to define A and B.

1.4.1 Rewriting A and B

$$
\begin{aligned}
B &= \mathrm{E}\left[uu^T\right] \\
&= \mathrm{E}\begin{bmatrix} (Y - \mu_1)Zw_1 \\ (Y - \mu_0)(1 - Z)w_0 \\ x(Z - e) \end{bmatrix} \begin{bmatrix} (Y - \mu_1)Zw_1, & (Y - \mu_0)(1 - Z)w_0, & x^T(Z - e) \end{bmatrix} \\
&= \mathrm{E}\begin{bmatrix} (Y - \mu_1)^2 Z w_1^2 & (Y - \mu_1)(Y - \mu_0)Z(1 - Z)w_1 w_0 & x^T(Y - \mu_1)Zw_1(Z - e) \\ (Y - \mu_1)(Y - \mu_0)Z(1 - Z)w_1 w_0 & (Y - \mu_0)^2(1 - Z)w_0^2 & x^T(Y - \mu_0)(1 - Z)(Z - e)w_0 \\ x(Z - e)(Y - \mu_1)Zw_1 & x(Z - e)(Y - \mu_0)(1 - Z)w_0 & xx^T(Z - e)^2 \end{bmatrix} \\
&= \begin{bmatrix} b_{11} & 0 & b_{13} \\ 0 & b_{22} & b_{23} \\ b_{13}^T & b_{23}^T & b_{33} \end{bmatrix} \text{ with}
\end{aligned}
$$

$$
\begin{aligned}
b_{11} &= \mathrm{E}\left[(Y - \mu_1)^2 Z w_1^2\right] \\
b_{13} &= \mathrm{E}\left[x^T(Y - \mu_1)Z(1 - e)w_1\right] \\
b_{22} &= \mathrm{E}\left[(Y - \mu_0)^2(1 - Z)w_0^2\right] \\
b_{23} &= -\mathrm{E}\left[x^T(Y - \mu_0)e(1 - Z)w_0\right] \\
b_{33} &= \mathrm{E}\left[xx^T(Z - e)^2\right],
\end{aligned}
$$

where b_{11}, b_{13}, b_{22}, b_{23}, b_{33} are scalar, $1 \times (p + 1)$ (row vector), scalar, $1 \times (p + 1)$ (row vector) and $(p + 1) \times (p + 1)$ matrix, respectively.

When replacing w_0 and w_1 with the ATE weights, we obtain an identical formula as in the appendix of Williamson's article, which is a specific case of the general formula. We rewrite A:

$$
\begin{aligned}
A &= -\mathrm{E}\left[\frac{\partial u}{\partial \theta^T}\right] = \\
&= -\mathrm{E}\begin{bmatrix} -Zw_1 & 0 & (Y - \mu_1)Z\frac{\partial w_1}{\partial \alpha^T} \\ 0 & -(1 - Z)w_0 & (Y - \mu_0)(1 - Z)\frac{\partial w_0}{\partial \alpha^T} \\ 0 & 0 & -xx^T e(1 - e) \end{bmatrix} = \begin{bmatrix} a_{11} & 0 & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}
\end{aligned}
$$

with:

$$
\begin{aligned}
a_{11} &= \mathrm{E}\left[Zw_1\right] \\
a_{13} &= -\mathrm{E}\left[(Y - \mu_1)Z\frac{\partial w_1}{\partial \alpha^T}\right] \\
a_{22} &= \mathrm{E}\left[(1 - Z)w_0\right] \\
a_{23} &= -\mathrm{E}\left[(Y - \mu_0)(1 - Z)\frac{\partial w_0}{\partial \alpha^T}\right] \\
a_{33} &= \mathrm{E}\left[xx^T e(1 - e)\right]
\end{aligned}
$$

where a_{11}, a_{13}, a_{22}, a_{23}, a_{33} are scalar, $1 \times (p+1)$ (row vector), scalar, $1 \times (p+1)$ (row vector) and $(p+1) \times (p+1)$ matrix, respectively.

1.4.2 Derivation of the final formula

Derivatives of the weights

For the different types of weights, we need $\frac{\partial w_i}{\partial \alpha^T}$. Note that these are functions of x_i and $\hat{\alpha}$:

ATE:

$$\frac{\partial w_1^{ATE}}{\partial \alpha^T} = -x^T \frac{1-e}{e} \quad \text{and} \quad \frac{\partial w_0^{ATE}}{\partial \alpha^T} = x^T \frac{e}{(1-e)}$$

ATT:

$$\frac{\partial w_1^{ATT}}{\partial \alpha^T} = 0 \quad \text{and} \quad \frac{\partial w_0^{ATT}}{\partial \alpha^T} = x^T \frac{e}{(1-e)}$$

Overlap weights:

$$\frac{\partial w_1^{OW}}{\partial \alpha^T} = -x^T e(1-e) \quad \text{and} \quad \frac{\partial w_0^{OW}}{\partial \alpha^T} = x^T e(1-e)$$

Matching weights [1]:

$$\frac{\partial w_1^{MW}}{\partial \alpha^T} = 0 \quad \text{if} \quad e_i < 0.5$$

$$\frac{\partial w_1^{MW}}{\partial \alpha^T} = -x^T \frac{1-e}{e} \quad \text{if} \quad e_i > 0.5$$

$$\frac{\partial w_0^{MW}}{\partial \alpha^T} = x^T \frac{e}{1-e} \quad \text{if} \quad e_i < 0.5$$

$$\frac{\partial w_0^{MW}}{\partial \alpha^T} = 0 \quad \text{if} \quad e_i > 0.5$$

Invert A

$AA^{-1} = \mathbb{1}$

By following simple linear algebra rules, we find:

$$A^{-1} = \begin{bmatrix} a_{11}^{-1} & 0 & -a_{11}^{-1}a_{13}a_{33}^{-1} \\ 0 & a_{22}^{-1} & -a_{22}^{-1}a_{23}a_{33}^{-1} \\ 0 & 0 & a_{33}^{-1} \end{bmatrix}$$

We can note that $a_{33} = b_{33}$. For ATE weights, there are more equivalences (see Appendix in Williamson *et al.*, 2014).

[1] Note that matching weights have a non-differentiable point at 0.5; see 'A weighting analogue to pair matching in propensity score analysis' and 'Comments on "A weighting analogue to pair matching in propensity score analysis" by L. Li and T. Greene'

Finally:

$$n^{-1}A^{-1}BA^{-T} = \begin{bmatrix} a_{11}^{-1} & 0 & -a_{11}^{-1}a_{13}a_{33}^{-1} \\ 0 & a_{22}^{-1} & -a_{22}^{-1}a_{23}a_{33}^{-1} \\ 0 & 0 & a_{33}^{-1} \end{bmatrix} \begin{bmatrix} b_{11} & 0 & b_{13} \\ 0 & b_{22} & b_{23} \\ b_{13}^T & b_{23}^T & b_{33} \end{bmatrix} \begin{bmatrix} a_{11}^{-1} & 0 & 0 \\ 0 & a_{22}^{-1} & 0 \\ -a_{11}^{-1}a_{33}^{-T}a_{13}^T & -a_{22}^{-1}a_{33}^{-T}a_{23}^T & a_{33}^{-T} \end{bmatrix} = \begin{bmatrix} Var(\mu_1) & Cov(\mu_1, \mu_0) & \times \\ \times & Var(\mu_0) & \times \\ \times & \times & \times \end{bmatrix}$$

(We replace a_{33}^{-T} with a_{33}^{-1}, because, the inverse of a symmetric matrix is also symmetric; in addition, elements not needed were replaced by $\times$); we end up with:

$$Var(\hat{\mu_1}) = n^{-1}a_{11}^{-2}\left[b_{11} - a_{13}a_{33}^{-1}b_{13}^T - b_{13}a_{33}^{-1}a_{13}^T + a_{13}a_{33}^{-1}b_{33}a_{33}^{-1}a_{13}^T\right]$$

$$Var(\hat{\mu_0}) = n^{-1}\left[a_{22}^{-2}\left[b_{22} - a_{23}a_{33}^{-1}b_{23}^T - b_{23}a_{33}^{-1}a_{23}^T + a_{23}a_{33}^{-1}b_{33}a_{33}^{-1}a_{23}^T\right]\right]$$

$$Cov(\hat{\mu_1}, \hat{\mu_0}) = n^{-1}\left[-a_{11}^{-1}a_{22}^{-1}\left[a_{13}a_{33}^{-1}b_{23}^T + b_{13}a_{33}^{-1}a_{23}^T - a_{13}a_{33}^{-1}b_{33}a_{33}^{-1}a_{23}^T\right]\right]$$

Large sample variance of $\hat{\delta}$

Then $Var(\hat{\delta}_j) = K_1^2 Var(\hat{\mu_1}) + K_0^2 Var(\hat{\mu_0}) - 2K_0K_1Cov(\hat{\mu_1}, \hat{\mu_0})$

The terms $K_.$ are specific to the measure of interest, δ, and are the derivatives of the link functions we would use for the estimation of these when fitting a classical GLM. More specifically:

- Risk or mean difference: $K_0 = K_1 = 1$ for $Z = 0$ and $Z = 1$, respectively.
- Risk ratio: $K_0 = \hat{\mu}_0^{-1}$ and $K_1 = \hat{\mu}_1^{-1}$ for $Z = 0$ and $Z = 1$, respectively.
- Odds ratio: $K_0 = [\hat{\mu}_0(1 - \hat{\mu}_0)]^{-1}$ and $K_1 = [\hat{\mu}_1(1 - \hat{\mu}_1)]^{-1}$ for $Z = 0$ and $Z = 1$, respectively.

Sample estimate of the variance of $\hat{\delta}$

Now, we can replace the components of A and B by their sample estimates:

$\hat{a}_{11} = n^{-1}\sum_{i=1}^{n} Z_i\hat{w}_{1i}$
$\hat{a}_{13} = -n^{-1}\sum_{i=1}^{n}(Y_i - \hat{\mu}_1)Z_i\frac{\partial \hat{w}_{1i}}{\partial \alpha^T}$
$\hat{a}_{22} = n^{-1}\sum_{i=1}^{n}(1 - Z_i)\hat{w}_{0i}$
$\hat{a}_{23} = -n^{-1}\sum_{i=1}^{n}(Y_i - \hat{\mu}_0)(1 - Z_i)\frac{\partial \hat{w}_{0i}}{\partial \alpha^T}$
$\hat{a}_{33} = n^{-1}\sum_{i=1}^{n}\hat{e}_i(1 - \hat{e}_i)x_ix_i^T$
$\hat{b}_{11} = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{\mu}_1)^2Z_i\hat{w}_{1i}^2$
$\hat{b}_{13} = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{\mu}_1)Z_i(1 - \hat{e}_i)\hat{w}_{1i}x_i^T$
$\hat{b}_{22} = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{\mu}_0)^2(1 - Z_i)\hat{w}_{0i}^2$
$\hat{b}_{33} = n^{-1}\sum_{i=1}^{n}(Z_i - \hat{e}_i)^2x_ix_i^T$
$\hat{b}_{23} = -n^{-1}\sum_{i=1}^{n}(Y_i - \hat{\mu}_0)\hat{e}_i(1 - Z_i)\hat{w}_{0i}x_i^T$

We can derive the sample estimate of the variance:

$$\begin{aligned}\hat{Var}(\hat{\delta}) = &\frac{\hat{K}_1^2}{n\hat{a}_{11}^2}\left[\hat{b}_{11} - \hat{a}_{13}\hat{a}_{33}^{-1}\hat{b}_{13}^T - \hat{b}_{13}\hat{a}_{33}^{-1}\hat{a}_{13}^T + \hat{a}_{13}\hat{a}_{33}^{-1}\hat{b}_{33}\hat{a}_{33}^{-1}\hat{a}_{13}^T\right] \\ &+ \frac{\hat{K}_0^2}{n\hat{a}_{22}^2}\left[\hat{b}_{22} - \hat{a}_{23}\hat{a}_{33}^{-1}\hat{b}_{23}^T - \hat{b}_{23}\hat{a}_{33}^{-1}\hat{a}_{23}^T + \hat{a}_{23}\hat{a}_{33}^{-1}\hat{b}_{33}\hat{a}_{33}^{-1}\hat{a}_{23}^T\right] \\ &+ \frac{2\hat{K}_0\hat{K}_1}{n\hat{a}_{11}\hat{a}_{22}}\left[\hat{a}_{13}\hat{a}_{33}^{-1}\hat{b}_{23}^T + \hat{b}_{13}\hat{a}_{33}^{-1}\hat{a}_{23}^T - \hat{a}_{13}\hat{a}_{33}^{-1}\hat{b}_{33}\hat{a}_{33}^{-1}\hat{a}_{23}^T\right]\end{aligned}$$

$\hat{Var}(\hat{\delta}) = Part\ 1 + Part\ 2 + Part\ 3$ with:

$$Part\ 1 = \frac{\hat{K}_1^2}{n\hat{a}_{11}^2}\hat{b}_{11} + \frac{\hat{K}_0^2}{n\hat{a}_{22}^2}\hat{b}_{22}$$

By denoting $\hat{\theta}_1 = \frac{\hat{K}_1}{\hat{a}_{11}}$ and $\hat{\theta}_0 = \frac{\hat{K}_0}{\hat{a}_{22}}$, we have:

$$Part\ 1 = \frac{1}{n}\hat{\theta}_1^2\hat{b}_{11} + \frac{1}{n}\hat{\theta}_0^2\hat{b}_{22} \tag{1.6}$$

Thus, Part 1 is $n^{-1}\hat{V}_{un}$. We move on to Part 2:

$$\begin{aligned}Part\ 2 = &\frac{\hat{K}_1^2}{n\hat{a}_{11}^2}\left[-\hat{a}_{13}\hat{a}_{33}^{-1}\hat{b}_{13}^T - \hat{b}_{13}\hat{a}_{33}^{-1}\hat{a}_{13}^T\right] \\ &+ \frac{\hat{K}_0^2}{n\hat{a}_{22}^2}\left[-\hat{a}_{23}\hat{a}_{33}^{-1}\hat{b}_{23}^T - \hat{b}_{23}\hat{a}_{33}^{-1}\hat{a}_{23}^T\right] \\ &+ \frac{2\hat{K}_0\hat{K}_1}{n\hat{a}_{11}\hat{a}_{22}}\left[\hat{a}_{13}\hat{a}_{33}^{-1}\hat{b}_{23}^T + \hat{b}_{13}\hat{a}_{33}^{-1}\hat{a}_{23}^T\right]\end{aligned}$$

We denote $\hat{M}_1 = \hat{a}_{33}^{-1}$:

$$\begin{aligned}Part\ 2 = &\frac{\hat{\theta}_1^2}{n}\left[-\hat{a}_{13}\hat{M}_1\hat{b}_{13}^T - \hat{b}_{13}\hat{M}_1\hat{a}_{13}^T\right] \\ &+ \frac{\hat{\theta}_0^2}{n}\left[-\hat{a}_{23}\hat{M}_1\hat{b}_{23}^T - \hat{b}_{23}\hat{M}_1\hat{a}_{23}^T\right] \\ &+ \frac{2\hat{\theta}_0\hat{\theta}_1}{n}\left[\hat{a}_{13}\hat{M}_1\hat{b}_{23}^T + \hat{b}_{13}\hat{M}_1\hat{a}_{23}^T\right]\end{aligned}$$

$$
\begin{aligned}
Part\ 2 = & -\frac{2\hat{\theta}_1^2}{n}\hat{a}_{13}\hat{M}_1\hat{b}_{13}^T \\
& -\frac{2\hat{\theta}_0^2}{n}\hat{a}_{23}\hat{M}_1\hat{b}_{23}^T \\
& +\frac{2\hat{\theta}_0\hat{\theta}_1}{n}\left(\hat{a}_{13}\hat{M}_1\hat{b}_{23}^T + \hat{b}_{13}\hat{M}_1\hat{a}_{23}^T\right)
\end{aligned} \tag{1.7}
$$

We define:

$$
\hat{v}_1 = -\hat{\theta}_1\hat{a}_{13}^T + \hat{\theta}_0\hat{a}_{23}^T \tag{1.8}
$$

and

$$
\hat{v}_2 = \hat{\theta}_1\hat{b}_{13}^T - \hat{\theta}_0\hat{b}_{23}^T \tag{1.9}
$$

Therefore:

$$
\begin{aligned}
2\hat{v}_1^T\hat{M}_1\hat{v}_2 &= (-2\hat{\theta}_1\hat{a}_{13}\hat{M}_1 + 2\hat{\theta}_0\hat{a}_{23}\hat{M}_1)(\hat{\theta}_1\hat{b}_{13}^T - \hat{\theta}_0\hat{b}_{23}^T) \\
&= -2\hat{\theta}_1^2\hat{a}_{13}\hat{M}_1\hat{b}_{13}^T + 2\hat{\theta}_0\hat{\theta}_1\hat{a}_{13}\hat{M}_1\hat{b}_{23}^T + 2\hat{\theta}_0\hat{\theta}_1\hat{a}_{23}\hat{M}_1\hat{b}_{13}^T - 2\hat{\theta}_0^2\hat{a}_{23}\hat{M}_1\hat{b}_{23}^T
\end{aligned}
$$

Noting that since $\hat{a}_{23}\hat{M}_1\hat{b}_{13}^T = \hat{b}_{13}\hat{M}_1\hat{a}_{23}^T$, (1.7) becomes:

$$
Part\ 2 = \frac{1}{n}\left(2\hat{v}_1^T\hat{M}_1\hat{v}_2\right)
$$

$$
Part\ 3 = \frac{\hat{K}_1^2}{n\hat{a}_{11}^2}\hat{a}_{13}\hat{a}_{33}^{-1}\hat{b}_{33}\hat{a}_{33}^{-1}\hat{a}_{13}^T + \frac{\hat{K}_0^2}{n\hat{a}_{22}^2}\hat{a}_{23}\hat{a}_{33}^{-1}\hat{b}_{33}\hat{a}_{33}^{-1}\hat{a}_{23}^T + \frac{2\hat{K}_0\hat{K}_1}{n\hat{a}_{11}\hat{a}_{22}}\hat{a}_{13}\hat{a}_{33}^{-1}\hat{b}_{33}\hat{a}_{33}^{-1}\hat{a}_{23}^T
$$

Denoting $\hat{M}_1 = \hat{a}_{33}^{-1}$ and $\hat{M}_2 = \hat{a}_{33}^{-1}\hat{b}_{33}\hat{a}_{33}^{-1}$, we have:

$$
\begin{aligned}
Part\ 3 &= \frac{\hat{K}_1^2}{n\hat{a}_{11}^2}\hat{a}_{13}\hat{M}_2\hat{a}_{13}^T + \frac{\hat{K}_0^2}{n\hat{a}_{22}^2}\hat{a}_{23}\hat{M}_2\hat{a}_{23}^T + \frac{2\hat{K}_0\hat{K}_1}{n\hat{a}_{11}\hat{a}_{22}}\hat{a}_{13}\hat{M}_2\hat{a}_{23}^T \\
&= \frac{\hat{\theta}_1^2}{n}\hat{a}_{13}\hat{M}_2\hat{a}_{13}^T + \frac{\hat{\theta}_0^2}{n}\hat{a}_{23}\hat{M}_2\hat{a}_{23}^T + \frac{2\hat{\theta}_0\hat{\theta}_1}{n}\hat{a}_{13}\hat{M}_2\hat{a}_{23}^T
\end{aligned}
$$

Replacing $\hat{v}_1$ equals to $-\hat{\theta}_1\hat{a}_{13}^T + \hat{\theta}_0\hat{a}_{23}^T$:

$$
\begin{aligned}
v_1^T \hat{M}_2 v_1 &= (-\hat{\theta}_1 \hat{a}_{13} \hat{M}_2 + \hat{\theta}_0 \hat{a}_{23} \hat{M}_2)(-\hat{\theta}_1 \hat{a}_{13}^T + \hat{\theta}_0 \hat{a}_{23}^T) \\
&= \hat{\theta}_1^2 \hat{a}_{13} \hat{M}_2 \hat{a}_{13}^T - \hat{\theta}_0 \hat{\theta}_1 \hat{a}_{13} \hat{M}_2 \hat{a}_{23}^T - \hat{\theta}_0 \hat{\theta}_1 \hat{a}_{23} \hat{M}_2 \hat{a}_{13}^T + \hat{\theta}_0^2 \hat{a}_{23} \hat{M}_2 \hat{a}_{23}^T
\end{aligned}
$$

$\hat{a}_{23} \hat{M}_2 \hat{a}_{13}^T = \hat{a}_{13} \hat{M}_2 \hat{a}_{23}^T$, so

$$
Part\ 3 = \frac{1}{n} v_1^T \hat{M}_2 v_1 \tag{1.10}
$$

Putting everything together, we have:

$$
\begin{aligned}
\hat{Var}(\hat{\delta}_j) &= \hat{K}_1^2 \hat{Var}(\hat{\mu}_1) + \hat{K}_0^2 \hat{Var}(\hat{\mu}_0) - 2\hat{K}_0 \hat{K}_1 \hat{Cov}(\hat{\mu}_1, \hat{\mu}_0) \\
&= Part\ 1 + Part\ 2 + Part\ 3 \\
&= \frac{1}{n} \left[\hat{V}_{un} + 2\hat{v}_1^T \hat{M}_1 \hat{v}_2 + v_1^T \hat{M}_2 v_1 \right]
\end{aligned} \tag{1.11}
$$

1.5 Equivalence of the two estimators

Since the estimated linearized variables are centered, the final variance estimator of the risk difference is approximately equal to:

$$
\hat{Var}_{lin}(\hat{\delta}_1) \approx n^{-2} \sum_{i=1}^{n} \left[\hat{I}_i(\hat{\mu}_1) - \hat{I}_i(\hat{\mu}_0) \right]^2 \tag{1.12}
$$

$$
\begin{aligned}
\hat{I}_i(\hat{\mu}_1) - \hat{I}_i(\hat{\mu}_0) &= \frac{1}{n^{-1} \sum_{j=1}^{n} \hat{w}_{1j} Z_j} \{\hat{w}_{1i} Z_i (Y_i - \hat{\mu}_1)\} \\
&+ \frac{1}{n^{-1} \sum_{j=1}^{n} \hat{w}_{1j} Z_j} \left[\frac{1}{n} \sum_{j=1}^{n} \hat{w}_{1j} Z_j (Y_j - \hat{\mu}_1) \right] \left[\frac{1}{n} \sum_{j=1}^{n} \hat{e}_j (1 - \hat{e}_j) x_j x_j \right]^{-1} (Z_i - \hat{e}_i) x_i \\
&- \frac{1}{n^{-1} \sum_{j=1}^{n} \hat{w}_{0j} (1 - Z_j)} \{\hat{w}_{0i} (1 - Z_i)(Y_i - \hat{\mu}_0)\} \\
&+ \frac{1}{n^{-1} \sum_{j=1}^{n} \hat{w}_{0j} (1 - Z_j)} \left[\frac{1}{n} \sum_{j=1}^{n} \hat{w}_{0j} (1 - Z_j)(Y_j - \hat{\mu}_0) \right] \left[\frac{1}{n} \sum_{j=1}^{n} \hat{e}_j (1 - \hat{e}_j) x_j x_j \right]^{-1} (Z_i - \hat{e}_i) x_i
\end{aligned} \tag{1.13}
$$

First, we can introduce some notations used for the development using M-estimation:

$$
\begin{aligned}
\hat{l}_i(\hat{\mu}_1) - \hat{l}_i(\hat{\mu}_0) &= \frac{1}{\hat{a}_{11}}\{\hat{w}_{1i} Z_i (Y_i - \hat{\mu}_1)\} \\
&+ \frac{1}{\hat{a}_{11}}\{-\hat{a}_{13}\}\, \hat{M}_1 (Z_i - \hat{e}_i) x_i \\
&- \frac{1}{\hat{a}_{22}}\{\hat{w}_{0i}(1 - Z_i)(Y_i - \hat{\mu}_0)\} \\
&+ \frac{1}{\hat{a}_{22}}\{-\hat{a}_{23}\}\, \hat{M}_1 (Z_i - \hat{e}_i) x_i \\
&= \hat{\theta}_1 \hat{w}_{1i} Z_i (Y_i - \hat{\mu}_1) - \hat{\theta}_0 \hat{w}_{0i}(1 - Z_i)(Y_i - \hat{\mu}_0) \\
&- (\hat{\theta}_1 \hat{a}_{13} - \hat{\theta}_0 \hat{a}_{23})\, \hat{M}_1 (Z_i - \hat{e}_i) x_i
\end{aligned} \tag{1.14}
$$

Then, we calculate:

$$
\begin{aligned}
(\hat{l}_i(\hat{\mu}_1) - \hat{l}_i(\hat{\mu}_0))^2 &= \hat{\theta}_1^2 \hat{w}_{1i}^2 Z_i^2 (Y_i - \hat{\mu}_1)^2 \qquad \text{term 1} = S_1 \\
&+ \hat{\theta}_0^2 \hat{w}_{0i}^2 (1 - Z_i)^2 (Y_i - \hat{\mu}_0)^2 \qquad \text{term 2} = S_2 \\
&+ [(\hat{\theta}_1 \hat{a}_{13} - \hat{\theta}_0 \hat{a}_{23})\, \hat{M}_1 (Z_i - \hat{e}_i) x_i]^2 \qquad \text{term 3} = S_3 \\
&- 2\hat{\theta}_1 \hat{w}_{1i} Z_i (Y_i - \hat{\mu}_1) \times \hat{\theta}_0 \hat{w}_{0i}(1 - Z_i)(Y_i - \hat{\mu}_0) \qquad \text{term 4} = S_4 \\
&- 2\hat{\theta}_1 \hat{w}_{1i} Z_i (Y_i - \hat{\mu}_1) \times (\hat{\theta}_1 \hat{a}_{13} - \hat{\theta}_0 \hat{a}_{23})\, \hat{M}_1 (Z_i - \hat{e}_i) x_i \qquad \text{term 5} = S_5 \\
&+ 2\hat{\theta}_0 \hat{w}_{0i}(1 - Z_i)(Y_i - \hat{\mu}_0) \times (\hat{\theta}_1 \hat{a}_{13} - \hat{\theta}_0 \hat{a}_{23})\, \hat{M}_1 (Z_i - \hat{e}_i) x_i \qquad \text{term 6} = S_6
\end{aligned} \tag{1.15}
$$

For term 1 and 2, we note that $T_i^2 = T_i$ and $(1 - T_i)^2 = 1 - T_i$, thus

$$
S_1 = \hat{\theta}_1^2 \hat{w}_{1i}^2 Z_i (Y_i - \hat{\mu}_1)^2 \tag{1.16}
$$

$$
S_2 = \hat{\theta}_0^2 \hat{w}_{0i}^2 (1 - Z_i)(Y_i - \hat{\mu}_0)^2 \tag{1.17}
$$

For term 3:

$$
\begin{aligned}
S_3 &= (\hat{\theta}_1 \hat{a}_{13} - \hat{\theta}_0 \hat{a}_{23})\, \hat{M}_1 (Z_i - \hat{e}_i)^2 x_i x_i\, \hat{M}_1 (\hat{\theta}_1 \hat{a}_{13} - \hat{\theta}_0 \hat{a}_{23}) \\
&= (-\hat{v}_1)\ \hat{M}_1 (Z_i - \hat{e}_i)^2 x_i x_i\, \hat{M}_1 (-\hat{v}_1) \\
&= \hat{v}_1\, \hat{M}_1 (Z_i - \hat{e}_i)^2 x_i x_i\, \hat{M}_1 \hat{v}_1
\end{aligned} \tag{1.18}
$$

Fort term 4: $Z_i(1 - Z_i) = 0$ thus

$$
S_4 = 0 \tag{1.19}
$$

For term 5: $Z_i(Z_i - \hat{e}_i) = Z_i(1 - \hat{e}_i)$ thus

$$S_5 = -2\hat{\theta}_1 \hat{w}_{1i} Z_i (1 - \hat{e}_i)(Y_i - \hat{\mu}_1) \times (-\hat{v}_1) \quad \hat{M}_1 x_i \tag{1.20}$$
$$= 2\hat{\theta}_1 \hat{w}_{1i} Z_i (1 - \hat{e}_i)(Y_i - \hat{\mu}_1)\hat{v}_1 \quad \hat{M}_1 x_i$$

For term 6: $(1 - Z_i)(Z_i - \hat{e}_i) = -(1 - Z_i)\hat{e}_i$ thus

$$S_6 = -2\hat{\theta}_0 \hat{w}_{0i} (1 - Z_i)\hat{e}_i (Y_i - \hat{\mu}_0)(-\hat{v}_1) \quad \hat{M}_1 x_i \tag{1.21}$$
$$= 2\hat{\theta}_0 \hat{w}_{0i} (1 - Z_i)\hat{e}_i (Y_i - \hat{\mu}_0)\hat{v}_1 \quad \hat{M}_1 x_i$$

Finally, we obtain:

$$\widehat{Var}_{lin}(\hat{\delta}_1) \approx n^{-2} \sum_{i=1}^{n} \left\{ \hat{I}_i(\hat{\mu}_1) - \hat{I}_i(\hat{\mu}_0) \right\}^2 \tag{1.22}$$
$$= \frac{1}{n} \left\{ \frac{\hat{\theta}_1^2}{n} \sum_{i=1}^{n} \hat{w}_{1i}^2 Z_i (Y_i - \hat{\mu}_1)^2 + \frac{\hat{\theta}_0^2}{n} \sum_{i=1}^{n} \hat{w}_{0i}^2 (1 - Z_i)(Y_i - \hat{\mu}_0)^2 \right.$$
$$+ \hat{v}_1 \hat{M}_1 \frac{1}{n} \sum_{i=1}^{n} (Z_i - \hat{e}_i)^2 x_i x_i \quad \hat{M}_1 \hat{v}_1$$
$$\left. + 2\hat{v}_1 \hat{M}_1 \left\{ \frac{\hat{\theta}_1}{n} \sum_{i=1}^{n} \hat{w}_{1i} Z_i (1 - \hat{e}_i)(Y_i - \hat{\mu}_1) x_i + \frac{\hat{\theta}_0}{n} \sum_{i=1}^{n} \hat{w}_{0i} (1 - Z_i)\hat{e}_i (Y_i - \hat{\mu}_0) x_i \right\} \right\}$$
$$= \frac{1}{n} \left\{ \hat{V}_{un} + \hat{v}_1 \hat{M}_2 \hat{v}_1 + 2\hat{v}_1 \hat{M}_1 \left\{ \hat{\theta}_1 \hat{b}_{13} - \hat{\theta}_0 \hat{b}_{23} \right\} \right\}$$
$$= \frac{1}{n} \left\{ \hat{V}_{un} + \hat{v}_1 \hat{M}_2 \hat{v}_1 + 2\hat{v}_1 \hat{M}_1 \hat{v}_2 \right\} = \widehat{Var}(\hat{\delta}_1)$$

In the same way, we obtain that:

$$\widehat{Var}_{lin}(\hat{\delta}_2) \approx \widehat{Var}(\hat{\delta}_2) \tag{1.23}$$
$$\widehat{Var}_{lin}(\hat{\delta}_3) \approx \widehat{Var}(\hat{\delta}_3) \tag{1.24}$$

1.6 Note: intermediate results for variance estimation via linearization

Herein, we present some intermediate algebra calculations to derive equations (21) - (23), (33), (40) and (41) demonstrated in section 4 of the main article:

1.6.1 Derivation of equation (21) in main manuscript

To derive $U_i(\alpha)$ we construct the score equation for α, following the generalized estimating equations theory [9]. Herein, we provide a short explanation on how to derive the term $\frac{\partial \{g^{-1}(x_i \ \alpha)\}}{\partial \alpha}$.

Let us write the individual propensity score as a function of the papameter vector, α: $e_i(\alpha) \equiv g^{-1}(x_i \ \alpha) : \alpha \quad R^{p+1} \rightarrow e_i(\alpha) \quad [0, 1]$. Let as also assume, a specific value, $\hat{\alpha} \quad R^{p+1}$, for which the propensity score takes the value $e_i = e_i(\hat{\alpha})$. The assumed link function for the propensity

score model is denoted as $g : e_i \quad [0,1] \rightarrow g(e_i) \quad R$:

$$\frac{\partial\{g^{-1}(x_i\ \alpha)\}}{\partial\alpha} \equiv \left[\frac{\partial\{g^{-1}(x_i\ \alpha)\}}{\partial\alpha_0}, \frac{\partial\{g^{-1}(x_i\ \alpha)\}}{\partial\alpha_1}, \ldots, \frac{\partial\{g^{-1}(x_i\ \alpha)\}}{\partial\alpha_p}\right] \Leftrightarrow \tag{1.25}$$

$$\frac{\partial\{g^{-1}(x_i\ \alpha)\}}{\partial\alpha} = \left[\frac{1}{\frac{\partial g}{\partial\alpha_0}}1, \frac{1}{\frac{\partial g}{\partial\alpha_1}}x_{1i}, \ldots, \frac{1}{\frac{\partial g}{\partial\alpha_p}}x_{pi}\right] = \left[\frac{1}{g\ (e_i)}1, \frac{1}{g\ (e_i)}x_{1i}, \ldots, \frac{1}{g\ (e_i)}x_{pi}\right] = \frac{1}{g\ (e_i)}x_i \tag{1.26}$$

To derive the second last equality in (1.26), we may apply the chain rule, separately to each component of the $(p + 1)$-row vector of the partial derivative of $e_i(\alpha)$ with respect to α. For example, to obtain the first element of the vector:

$$g(g^{-1}(x_i\ \alpha)) = x_i\ \alpha \Leftrightarrow \tag{1.27}$$

$$\frac{\partial[g(g^{-1}(x_i\ \alpha))]}{\partial\alpha_0} = 1 \Leftrightarrow \tag{1.28}$$

$$\frac{\partial g}{\partial\alpha_0} \times \frac{\partial g^{-1}(x_i\ \alpha)}{\partial\alpha_0} = 1 \Leftrightarrow \tag{1.29}$$

$$\frac{\partial g^{-1}(x_i\ \alpha)}{\partial\alpha_0} = \frac{1}{\frac{\partial g}{\partial\alpha_0}} \tag{1.30}$$

Likewise, for the second element, by applying the chain rule, we have:

$$g(g^{-1}(x_i\ \alpha)) = x_i\ \alpha \Leftrightarrow \tag{1.31}$$

$$\frac{\partial[g(g^{-1}(x_i\ \alpha))]}{\partial\alpha_1} = x_{1i} \Leftrightarrow \tag{1.32}$$

$$\frac{\partial g}{\partial\alpha_1} \times \frac{\partial g^{-1}(x_i\ \alpha)}{\partial\alpha_1} = x_{1i} \Leftrightarrow \tag{1.33}$$

$$\frac{\partial g^{-1}(x_i\ \alpha)}{\partial\alpha_1} = \frac{1}{\frac{\partial g}{\partial\alpha_1}}x_{1i}, \tag{1.34}$$

and so on...

1.6.2 Derivation of equation (22) in main manuscript

If $\bar{U}(\cdot)$ is continuously differentiable in a neighborhood of α, and from the mean value theorem, there exists some vector $\tilde{\alpha}$ whose components lie between those of α and $\hat{\alpha}$ and such that

$$\begin{aligned}\bar{U}(\hat{\alpha}) - \bar{U}(\alpha) &= \left[\frac{\partial \bar{U}}{\partial \alpha}(\tilde{\alpha})\right](\hat{\alpha} - \alpha) \\ &= E\left[\frac{\partial U(\alpha)}{\partial \alpha}\right](\hat{\alpha} - \alpha) + \left[\frac{\partial \bar{U}}{\partial \alpha}(\tilde{\alpha}) - E\left[\frac{\partial U(\alpha)}{\partial \alpha}\right]\right](\hat{\alpha} - \alpha). \qquad (1.35)\end{aligned}$$

Under the Mean Value Theorem conditions (note: if $U_i(\alpha)$ is a $p + 1$ column vector, the key is to consider each of the $p + 1$ components separately; briefly, for the mean value theorem to be applicable, each of the $p + 1$ components of $U_i(\alpha)$ must be continuous on the closed $[\alpha, \hat{\alpha}]$ and differentiable on the open interval; for more consult Boos and Stefanski: "Essential statistical inference: theory and methods"), the second term on the right-hand side of (1.35) is $o_p(n^{-1/2})$, which leads to eq. (22) in the main manuscript.

1.6.3 Derivation of equation (23) in main manuscript

From (21), we have

$$U_i(\alpha) = \frac{f_{1i}(\alpha)}{f_{2i}(\alpha)}, \qquad (1.36)$$

with $f_{1i}(\alpha) = (Z_i - e_i)x_i$ and $f_{2i}(\alpha) = e_i(1 - e_i)g(e_i)$. Since from (1.26)

$$\frac{\partial e_i}{\partial \alpha} = \frac{x_i}{g(e_i)},$$

we obtain

$$\begin{aligned}\frac{\partial f_{1i}(\alpha)}{\partial \alpha} &= -\frac{x_i x_i}{g(e_i)}, \qquad (1.37)\\ \frac{\partial f_{2i}(\alpha)}{\partial \alpha} &= \left[\frac{x_i}{g(e_i)} - 2e_i\frac{x_i}{g(e_i)}\right] g(e_i) + e_i(1 - e_i)\frac{\partial g(e_i)}{\partial \alpha} \\ &= x_i(1 - 2e_i) + e_i(1 - e_i)g(e_i) \times \frac{x_i}{g(e_i)}.\end{aligned}$$

From (1.36), we have

$$\begin{aligned}\frac{\partial U_i(\alpha)}{\partial \alpha} &= \frac{1}{[f_{2i}(\alpha)]^2}\left[f_{2i}(\alpha)\frac{\partial f_{1i}(\alpha)}{\partial \alpha} - f_{1i}(\alpha)\frac{\partial f_{2i}(\alpha)}{\partial \alpha}\right] \\ &= \frac{1}{[e_i(1 - e_i)g(e_i)]^2}\left[-e_i(1 - e_i)g(e_i) \times \frac{x_i x_i}{g(e_i)} - (Z_i - e_i)x_i\left[x_i(1 - 2e_i) + e_i(1 - e_i)x_i\frac{g(e_i)}{g(e_i)}\right]\right] \\ &= -\frac{x_i x_i}{e_i(1 - e_i)[g(e_i)]^2} - (Z_i - e_i)x_i x_i\frac{(1 - 2e_i) + e_i(1 - e_i)\frac{g(e_i)}{g(e_i)}}{[e_i(1 - e_i)g(e_i)]^2}. \qquad (1.38)\end{aligned}$$

Conditionally on x_i, the expectation of the second term in the right-hand side of (1.38) is equal to zero. Therefore, the unconditional expectation of this term is also equal to zero, and we obtain (23).

1.6.4 Derivation of equation (33) in main manuscript

We have

$$\begin{aligned}\hat{\mu}_1 - \mu_1 &= \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{w}_{1i}Z_i(Y_i-\mu_1)}{\frac{1}{n}\sum_{i=1}^{n}\hat{w}_{1i}Z_i} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{w}_{1i}Z_i(Y_i-\mu_1)}{E(w_1Z)} \times \left(1 + \frac{E(w_1Z) - \frac{1}{n}\sum_{i=1}^{n}\hat{w}_{1i}Z_i}{\frac{1}{n}\sum_{i=1}^{n}\hat{w}_{1i}Z_i}\right) \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{w}_{1i}Z_i(Y_i-\mu_1)}{E(w_1Z)} \times \left(1 + O_p(n^{-1/2})\right), \end{aligned} \tag{1.39}$$

which leads to (33).

1.6.5 Derivation of equations (41) and (42) in main manuscript

We have $\delta_2 = \log(\mu_1) - \log(\mu_0)$. By using Rule 1 in [7] for the logarithmic function, we have

$$I_i(\delta_2) = \frac{I_i(\mu_1)}{\mu_1} - \frac{I_i(\mu_0)}{\mu_0}, \tag{1.40}$$

which leads to (41) after replacing unknown quantities by empirical estimators. Also, we have

$$\delta_3 = \log(\mu_1) - \log(1-\mu_1) - \log(\mu_0) + \log(1-\mu_0),$$

and using the same rule we have

$$\begin{aligned} I_i(\delta_3) &= \frac{I_i(\mu_1)}{\mu_1} + \frac{I_i(\mu_1)}{1-\mu_1} - \frac{I_i(\mu_0)}{\mu_0} - \frac{I_i(\mu_0)}{1-\mu_0} \\ &= \frac{I_i(\mu_1)}{\mu_1(1-\mu_1)} - \frac{I_i(\mu_0)}{\mu_0(1-\mu_0)}, \end{aligned}$$

which leads to (42) after replacing unknown quantities by empirical estimators.

2 NSCLC data: bootstrap application

In this section, we note that we may obtain asymptotically equivalent results to the analytic IPTW large-sample variance estimate by applying the non-parametric bootstrap technique. The bootstrap, just like the analytic method we followed, provides valid results under standard assumptions (i.e., as long as the sample size is sufficiently large - which is also required for the analytic estimator to be valid) [3]. We note that we need the bootstrap principle to hold [4]. For IPTW, this can be problematic when very few individuals have large weights (because, whether they were selected or not in the bootstrap sample would make a big difference). Within each bootstrap sample, we need to re-estimate the individual PS, as the aim is to balance the covariate distributions between treatment groups in the analysis sample. For the bootstrap 95% confidence intervals we use the normal approximation bootstrap CI, where the correct standard error is the empirical standard deviation of the bootstrap replicates. Despite the ease in application, the bootstrap tends to be time-consuming for the case of big sets of epidemiological data in comparison to the analytic estimator. Nonetheless, it is frequently used to calculate correctly the required confidence intervals

for the IPTW estimates at the nominal level. In boxes 9a and 9b of the code Appendix we demonstrate the R code corresponding to the bootstrap application for different weighting schemes. In our example we used $R = 1,000$ bootstrap replicates.

3 Packages geex and PSweight in R: an alternative to manual implementation

To account for the uncertainty in the PS estimates when estimating the variance of the point estimates, one could apply either the R package geex [5] or PSweight [6]. The example code is demonstrated in the code Appendix.

4 Simulation study

4.1 Additional results

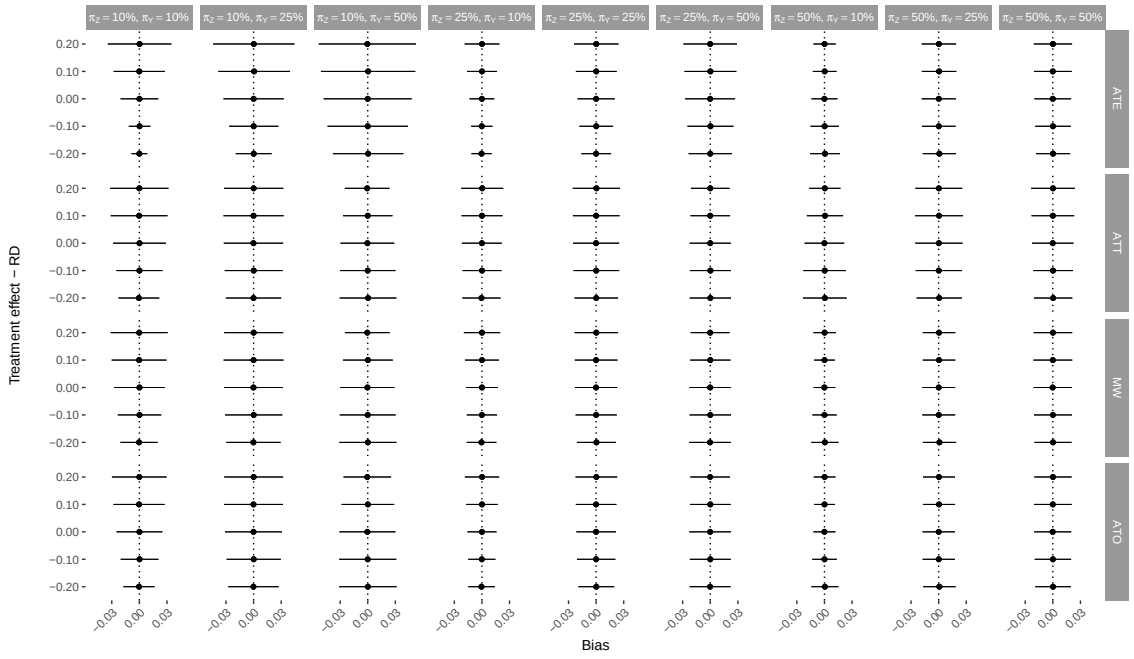


Figure 1: Bias in the risk difference estimates for simulation studies (n_{sim}=10,000) for four different Propensity Score (PS) weighting schemes: Average Treatment Effect (ATE), Average Treatment Effect in the Treated (ATT), Matching Weights (MW) and Overlap Weights (OW), and nine different combinations of (true) marginal treatment prevalences (π_Z) and outcome rates (π_Y)

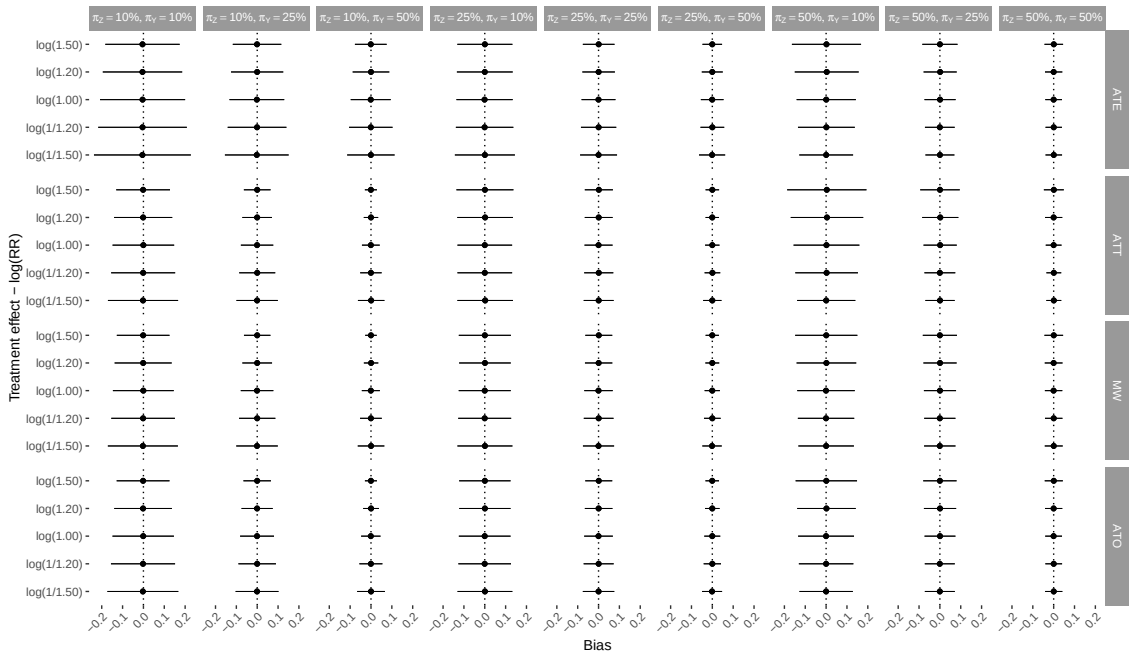


Figure 2: Bias in the logarithm of the marginal risk ratio estimates for simulation studies (n_{sim}=10,000) for four di□erent Propensity Score (PS) weighting schemes: Average Treatment E□ect (ATE), Average Treatment E□ect in the Treated (ATT), Matching Weights (MW) and Overlap Weights (OW), and nine di□erent combinations of (true) marginal treatment prevalences (π_Z) and outcome rates (π_Y)

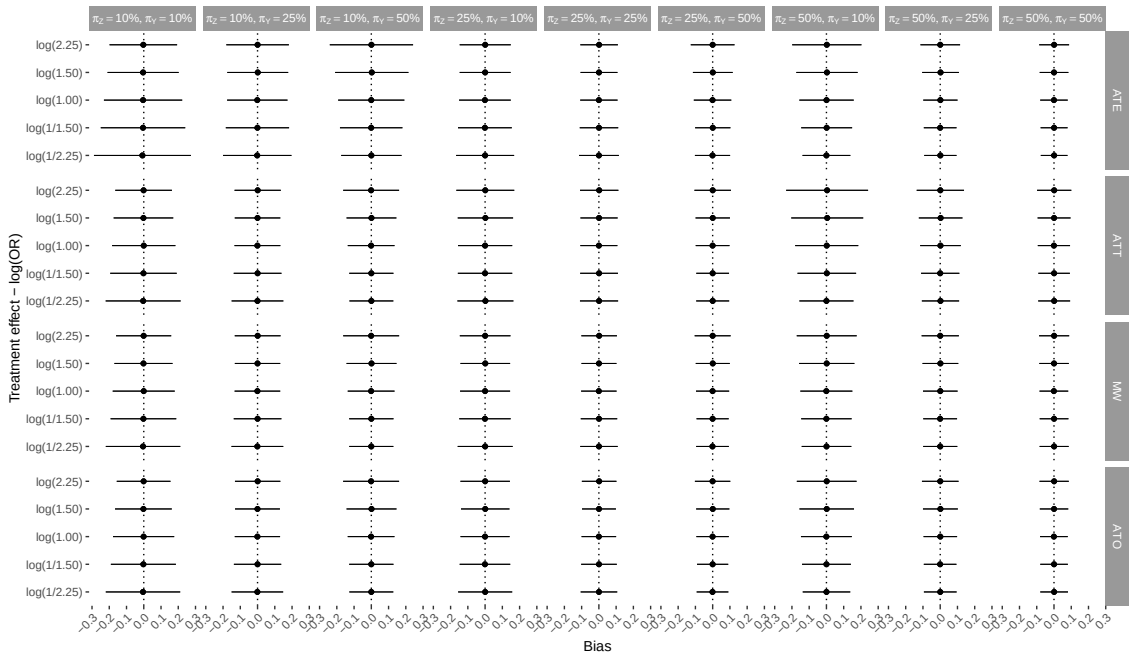


Figure 3: Bias in the logarithm of the marginal odds ratio estimates for simulation studies (n_{sim}=10,000) for four di□erent Propensity Score (PS) weighting schemes: Average Treatment E□ect (ATE), Average Treatment E□ect in the Treated (ATT), Matching Weights (MW) and Overlap Weights (OW), and nine di□erent combinations of (true) marginal treatment prevalences (π_Z) and outcome rates (π_Y)

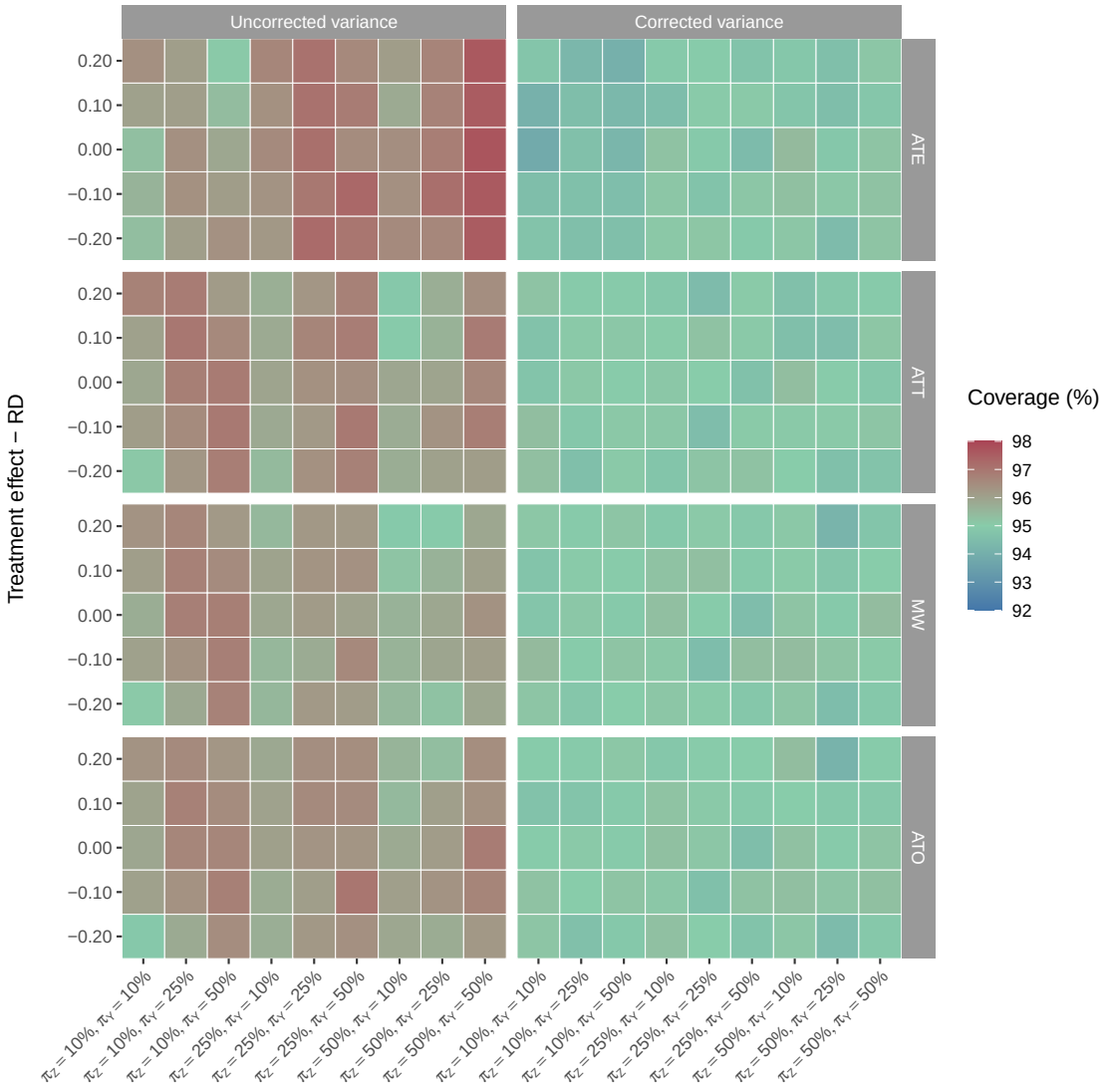


Figure 4: Coverage rate of 95% confidence intervals in simulation studies (n_{sim} =10,000) comparing the corrected *vs* uncorrected variance estimator for the Risk Di□erence (RD) for four di□erent Propensity Score (PS) weighting schemes: Average Treatment E□ect (ATE), Average Treatment E□ect in the Treated (ATT), Matching Weights (MW) and Overlap Weights (OW)

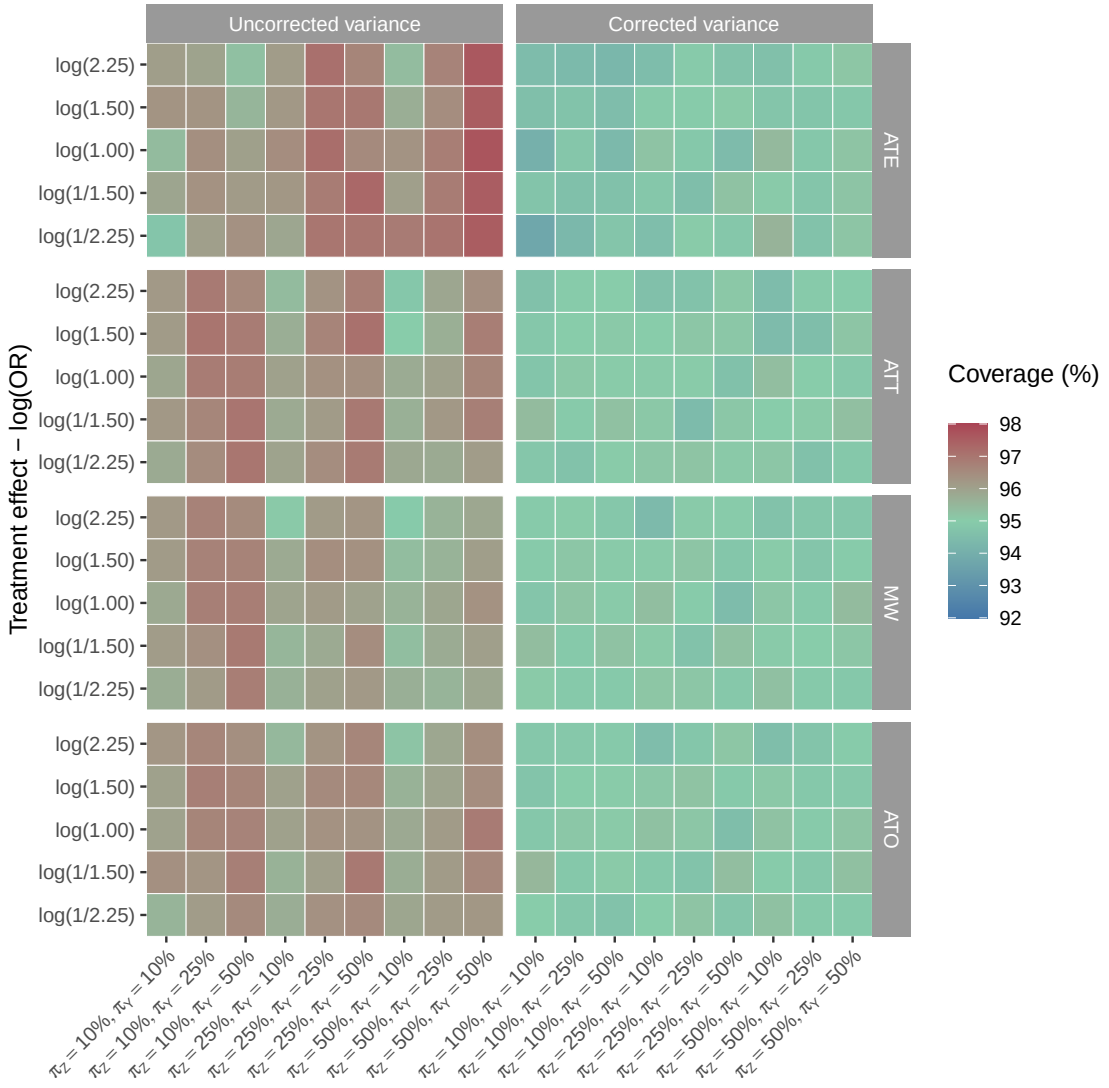


Figure 5: Coverage rate of 95% confidence intervals in simulation studies (n_{sim} =10,000) comparing the corrected *vs* uncorrected variance estimator for the log Odds Ratio (OR) for four di□erent Propensity Score (PS) weighting schemes: Average Treatment E□ect (ATE), Average Treatment E□ect in the Treated (ATT), Matching Weights (MW) and Overlap Weights (OW)

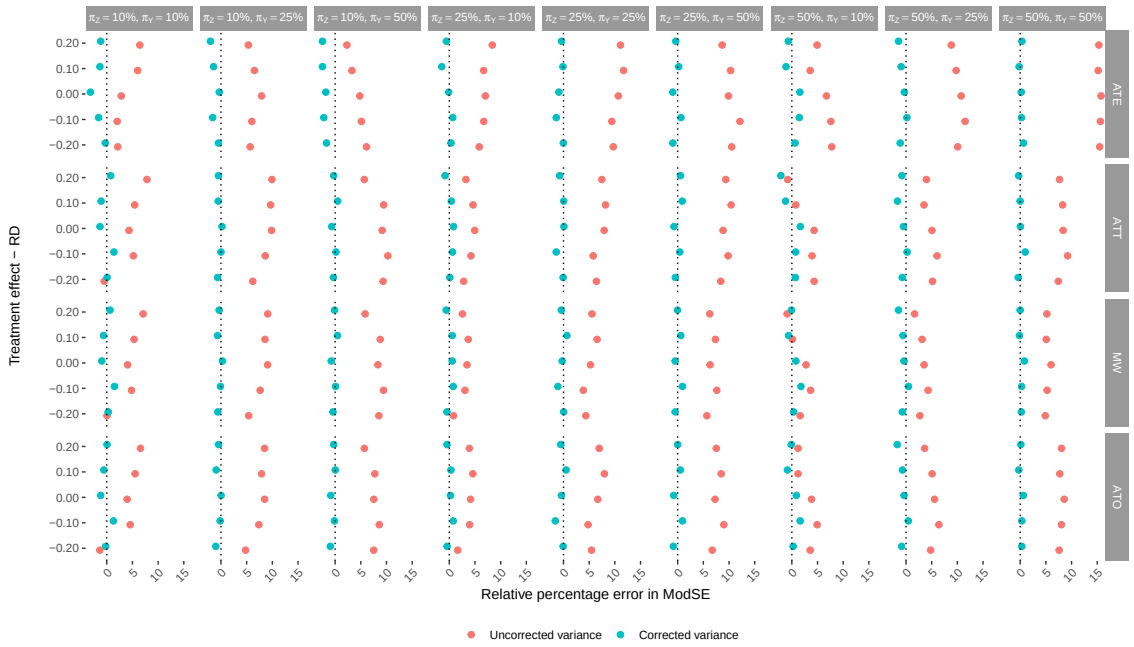


Figure 6: Relative Error (%) in Model Based Standard Error in simulation studies (n_{sim}=10,000) comparing the corrected *vs* uncorrected variance estimator for the Risk Di□erence (RD) for four di□erent Propensity Score (PS) weighting schemes: Average Treatment E□ect (ATE), Average Treatment E□ect in the Treated (ATT), Matching Weights (MW) and Overlap Weights (OW)

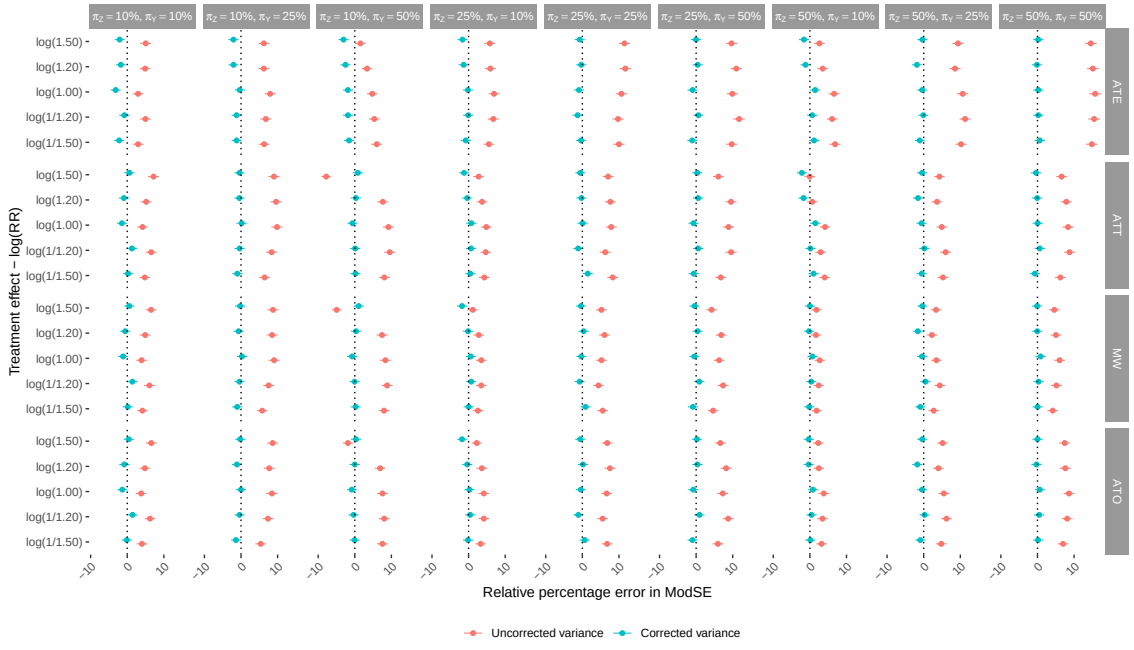


Figure 7: Relative Error (%) in Model Based Standard Error in simulation studies (n_{sim}=10,000) comparing the corrected *vs* uncorrected variance estimator for the log Risk Ratio (RR) for four di□erent Propensity Score (PS) weighting schemes: Average Treatment E□ect (ATE), Average Treatment E□ect in the Treated (ATT), Matching Weights (MW) and Overlap Weights (OW)

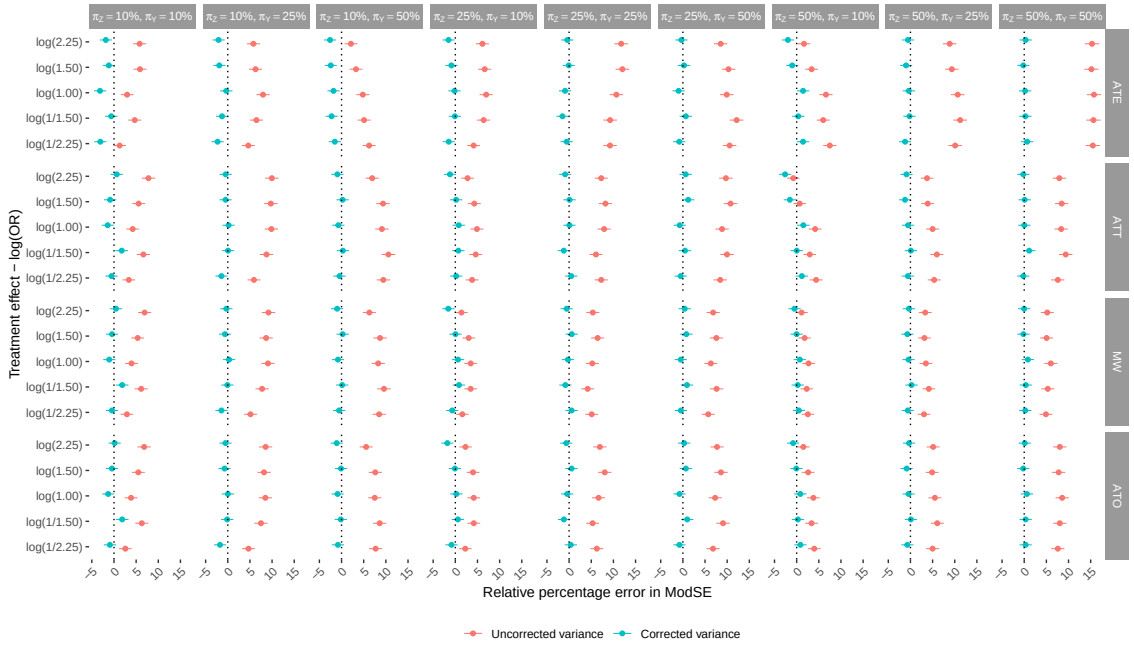


Figure 8: Relative Error (%) in Model Based Standard Error in simulation studies (n_{sim} =10,000) comparing the corrected *vs* uncorrected variance estimator for the log Odds Ratio (OR) for four di□erent Propensity Score (PS) weighting schemes: Average Treatment E□ect (ATE), Average Treatment E□ect in the Treated (ATT), Matching Weights (MW) and Overlap Weights (OW)

4.2 Simulation parameters

We used a data-generating process similar to the one found in Austin *et al.* studies to examine different aspects of propensity score analysis [1], [2].

First, we randomly generated 10 independent normally distributed ($N(0,1)$) variables $X_1 \dots X_{10}$ for n=10,000 subjects. The exposure allocation Z was drawn from a Bernoulli distribution $Z \sim B(p_Z)$, with

$$\begin{aligned} p_Z &= \text{logit}^{-1}(\alpha_{0,Z} \\ &\quad + \alpha_L X_1 + \alpha_L X_2 + \alpha_L X_3 \\ &\quad + \alpha_M X_4 + \alpha_M X_5 + \alpha_M X_6 \\ &\quad + \alpha_H X_7 + \alpha_H X_8 + \alpha_H X_9 + \alpha_{VH} X_{10}). \end{aligned} \tag{4.1}$$

A binary event was also generated for each subject, with a probability p_Y equal to

$$\begin{aligned} p_Y &= \text{logit}^{-1}(\alpha_{0,Y} + \gamma Z \\ &\quad + \alpha_L X_1 + \alpha_L X_2 + \alpha_L X_3 \\ &\quad + \alpha_M X_4 + \alpha_M X_5 + \alpha_M X_6 \\ &\quad + \alpha_H X_7 + \alpha_H X_8 + \alpha_H X_9 + \alpha_{VH} X_{10}). \end{aligned} \tag{4.2}$$

In the previous equation, γ denotes the conditional log odds ratio relating the treatment Z to the outcome Y. The other regression coefficients were set as follows to reflect low, medium, high and very high effects: $\alpha_L = \log(1.1)$, $\alpha_M = \log(1.25)$, $\alpha_H = \log(1.5)$ and $\alpha_{VH} = \log(2)$.

$\alpha_{0,Z}$, $\alpha_{0,Y}$ and γ were set to values that induce the desired treatment prevalence π_Z, event rate π_Y and marginal effect Γ (RD, RR, or OR, ATE, ATT, MW, or ATO estimators) in the simulated sample. These three parameters are mutually dependent, and we used an iterative process to determine the values of $\alpha_{0,Z}$, $\alpha_{0,Y}$, and γ that induce desired π_Z, π_Y and Γ. First, we simulated n=10,000 subjects, and computed the individual probabilities of being exposed ($\tilde{p}_{Z,i}$) with equation (4.1). The average of these individual probabilities is the expected exposure prevalence $\tilde{\pi}_Z = \frac{1}{n}\sum_{i=1}^{n}\tilde{p}_{Z,i}$ in the simulated sample. Similarly, we computed the individual probabilities of event ($\tilde{p}_{Y,i}$) with equation (4.2), and the corresponding average, which is the expected event rate $\tilde{\pi}_Y = \frac{1}{n}\sum_{i=1}^{n}\tilde{p}_{Y,i}$ in the sample.

We also computed the average probability of event first assuming that all subjects were untreated ($\tilde{\pi}_{Y,0} = \frac{1}{n}\sum_{i=1}^{n}\tilde{p}_{Y_0,i}$) and then assuming that all subjects were treated ($\tilde{\pi}_{Y,1} = \frac{1}{n}\sum_{i=1}^{n}\tilde{p}_{Y_1,i}$). The difference between these two average probabilities is the expected risk difference in the overall population, $\tilde{\Gamma}_{1,RD} = \tilde{\pi}_{Y,1} - \tilde{\pi}_{Y,0}$, in the sample. We computed the same average probabilities weighted by individual probabilities of being exposed ($\tilde{\pi}_{Y,0} = \frac{1}{n}\sum_{i=1}^{n}\tilde{p}_{Z,i}\tilde{p}_{Y_0,i}$ and $\tilde{\pi}_{Y,1} = \frac{1}{n}\sum_{i=1}^{n}\tilde{p}_{Z,i}\tilde{p}_{Y_1,i}$). The difference between these two weighted average probabilities is the expected risk difference in the treated population, $\tilde{\Gamma}_{2,RD} = \tilde{\pi}_{Y,1} - \tilde{\pi}_{Y,0}$, in the sample. We computed the same average probabilities weighted by $\min(\tilde{p}_{Z,i}, 1-\tilde{p}_{Z,i})$ ($\tilde{\pi}_{Y,0} = \frac{1}{n}\sum_{i=1}^{n}\min(\tilde{p}_{Z,i}, 1-\tilde{p}_{Z,i})\tilde{p}_{Y_0,i}$ and $\tilde{\pi}_{Y,1} = \frac{1}{n}\sum_{i=1}^{n}\min(\tilde{p}_{Z,i}, 1-\tilde{p}_{Z,i})\tilde{p}_{Y_1,i}$). The difference between these two weighted average probabilities is the expected risk difference in the 'matching weights' population, $\tilde{\Gamma}_{3,RD} = \tilde{\pi}_{Y,1} - \tilde{\pi}_{Y,0}$,

in the sample. Finally, we computed the same average probabilities weighted by $\tilde{p}_{Z,i}(1-\tilde{p}_{Z,i})$ ($\tilde{\pi}_{Y,0} = \frac{1}{n}\sum_{i=1}^{n}\tilde{p}_{Z,i}(1-\tilde{p}_{Z,i})\tilde{p}_{Y_0,i}$ and $\tilde{\pi}_{Y,1} = \frac{1}{n}\sum_{i=1}^{n}\tilde{p}_{Z,i}(1-\tilde{p}_{Z,i})\tilde{p}_{Y_1,i}$). The difference between these two weighted average probabilities is the expected risk difference in the 'overlap' population, $\tilde{\Gamma}_{4,RD} = \tilde{\pi}_{Y,1} - \tilde{\pi}_{Y,0}$, in the sample.

Using an iterative process, one could successively modify $\alpha_{0,Z}$, $\alpha_{0,Y}$ and γ until the expected treatment prevalence, the expected event rate and the expected marginal risk difference are arbitrarily close to the desired value in the simulated cohort. This process was performed by minimizing:

- the quantity $(\pi_Z - \tilde{\pi}_Z)^2 + (\pi_E - \tilde{\pi}_E)^2 + (\Gamma_{RD} - \tilde{\Gamma}_{1,RD})^2$ to obtain the parameters $\alpha_{0,Z}$, $\alpha_{0,Y}$ and γ that induced the desired exposure prevalence, event rate, and risk difference in the overall population (ATE);
- the quantity $(\pi_Z - \tilde{\pi}_Z)^2 + (\pi_E - \tilde{\pi}_E)^2 + (\Gamma_{RD} - \tilde{\Gamma}_{2,RD})^2$ to obtain the parameters $\alpha_{0,Z}$, $\alpha_{0,Y}$ and γ that induced the desired exposure prevalence, event rate, and risk difference in the treated population (ATT);
- the quantity $(\pi_Z - \tilde{\pi}_Z)^2 + (\pi_E - \tilde{\pi}_E)^2 + (\Gamma_{RD} - \tilde{\Gamma}_{3,RD})^2$ to obtain the parameters $\alpha_{0,Z}$, $\alpha_{0,Y}$ and γ that induced the desired exposure prevalence, event rate, and risk difference in the 'matching weights' population (MW);
- and the quantity $(\pi_Z - \tilde{\pi}_Z)^2 + (\pi_E - \tilde{\pi}_E)^2 + (\Gamma_{RD} - \tilde{\Gamma}_{4RD})^2$ to obtain the parameters $\alpha_{0,Z}$, $\alpha_{0,Y}$ and γ that induced the desired exposure prevalence, event rate, and risk difference in the 'overlap' population (ATO).

To increase precision, this minimization process was repeated in 1,000 simulated samples, to obtain 1000 sets of parameters $\alpha_{0,Z}$, $\alpha_{0,Y}$ and γ for ATE, ATT, MW and ATO risk differences. These 1000 estimations were averaged to obtain the final parameters used in the simulation study.

The parameters suitable for relative risks and odds ratios were obtained using a similar approach, replacing $\tilde{\Gamma}_{1,RD}$, $\tilde{\Gamma}_{2,RD}$, $\tilde{\Gamma}_{3,RD}$ and $\tilde{\Gamma}_{4,RD}$:

- by $\tilde{\Gamma}_{1,RR} = \log(\tilde{\pi}_{Y,1}) - \log(\tilde{\pi}_{Y,0})$, $\tilde{\Gamma}_{2,RR} = \log(\tilde{\pi}_{Y,1}) - \log(\tilde{\pi}_{Y,0})$, $\tilde{\Gamma}_{3,RR} = \log(\tilde{\pi}_{Y,1}) - \log(\tilde{\pi}_{Y,0})$ and $\tilde{\Gamma}_{4,RR} = \log(\tilde{\pi}_{Y,1}) - \log(\tilde{\pi}_{Y,0})$;
- or by $\tilde{\Gamma}_{1,OR} = \text{logit}(\tilde{\pi}_{Y,1}) - \text{logit}(\tilde{\pi}_{Y,0})$, $\tilde{\Gamma}_{2,OR} = \text{logit}(\tilde{\pi}_{Y,1}) - \text{logit}(\tilde{\pi}_{Y,0})$, $\tilde{\Gamma}_{3,OR} = \text{logit}(\tilde{\pi}_{Y,1}) - \text{logit}(\tilde{\pi}_{Y,0})$ and $\tilde{\Gamma}_{4,OR} = \text{logit}(\tilde{\pi}_{Y,1}) - \text{logit}(\tilde{\pi}_{Y,0})$.

One can notice (with equation 4.1) that the probability of treatment depends on only the subjects characteristics. Thus, for a given desired treatment prevalence, all the parameters $\alpha_{0,Z}$ obtained with the previously described minimization process were approximatively equal, whatever the desired event rate and treatment effect. Consequently, $\alpha_{0,Z}$ was considered unique for a given treatment prevalence (i.e. the values obtained for the same value of treatment prevalence were averaged).

5 Glossary

Below we provide brief explanations on some technical terms mentioned in the main manuscript:

- **Delta method**: for the (simplest) scalar case, we may write: assume the estimator of a true parameter, $\hat{\theta}$, follows an asymptotically normal distribution, i.e., $\hat{\theta} \quad AN(\theta, \sigma_n^2)$ with $\sigma_n \to \infty$ and g a real-valued function differentiable at θ, with $g'(\theta) = 0$. Then, as $n \to \infty$, $g(\hat{\theta}) \quad AN(g(\theta), \{g'(\theta)\}^2\sigma_n^2)$. For versions of the Delta method in the multivariate case - i.e., when θ and/or g are vectors rather than scalars, consult Chapter 1 (p. 14) in [8].

- **M-estimation**: a method used to estimate a target parameter, θ, by an estimator, $\hat{\theta}$; in particular, an M-estimator is any solution for θ, in $\sum_{i=1}^{n} \psi(O_i; \hat{\theta}) = 0$, where O_i are independent observations, θ, is a $k-$vector of parameters to estimate, $\Psi(.)$ is a known $k-$dimensional function, which does not depend on either i or n; a proof of consistency and asymptotic normality of the estimator can be demonstrated (e.g., see pp. 327-329 in [8]).

- **Taylor expansion**: assume that $\mathbf{g} : R^s \to R^k$. Irrespective of the dimensions s and k, the derivative of $\mathbf{g}$ at a particular point $\mathbf{x}_0$ offers a linear approximation to $\mathbf{g}(\mathbf{x})$ for *"$\mathbf{x}$ close to $\mathbf{x}_0$"*, which is of the form: $\underset{k\times 1}{\mathbf{g}(\mathbf{x})} \approx \underset{k\times 1}{\mathbf{g}(\mathbf{x}_0)} + \underset{k\times s}{\mathbf{g}'(\mathbf{x}_0)} \underset{s\times 1}{(\mathbf{x} - \mathbf{x}_0)}$, and is the best, meaning that the derivative $\mathbf{g}'$ has the defining property:

$$\lim_{\delta \to 0} \frac{||\mathbf{g}(\mathbf{x} + \delta) - \{\mathbf{g}(\mathbf{x}) + \mathbf{g}'(\mathbf{x})\delta\}||}{||\delta||} = 0$$

  where, $||.||$, the Euclidean norm, i.e., $||\mathbf{u}_{n\times 1}|| = (\mathbf{u}^T\mathbf{u})^{1/2}$. For more, consult [8].

References

[1] Peter C. Austin. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine*, 29(20):2137–2148, 2010.

[2] Peter C. Austin. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics: the journal of the pharmaceutical industry*, 10(2):150–161, 2011.

[3] James Carpenter and Bithell John. Bootstrap confidence intervals: When, which, what? a practical guide for medical statisticians. *Statistics in Medicine*, 19(9):1141–1164, 2000.

[4] B. Efron and R.J. Tibshirani. *An introduction to the Bootstrap*. Chapman and Hall, 1993.

[5] Bradley C. Saul and Michael G. Hudgens. The calculus of m-estimation in r with geex. *Journal of statistical software*, 92(2):1–15, 2020.

[6] T. Zhou, G. Tong, F. Li, LE. Thomas, and F. Li. Psweight: an r package for propensity score weighting analysis. *arXiv:2010.08893v4*, 2021.

[7] J. Deville Variance estimation for complex statistics and estimators: Linearization and residual techniques Survey Methodology, 25(2):193–203, 1999.

[8] Dennis D. Boos and, Leonard A. Stefanski. Essential Statistical Inference: Theory and Methods Springer, 2013.

[9] James W. Hardin and, Joseph M. Hilbe. Generalized Estimating Equations Chapman and Hall/CRC, 2002.

Appendix C

Additional figures for bias results of simulation study in section 4.4

Below, we present the nested loop plots [60] for a null $p \quad ATE$[1,2]. On the y axis we have the bias values. On the x axis, we have the corresponding ordered DGM indices (with index $2\ f1;2;:::;288g$; a detailed description of each DGM code can be found in the corresponding .csv file located in the GitHub repository sims-ISCB2023). Different colors correspond to different analysis methods[3]. Methods with bias values closer to zero were the least biased.

In figure C.1, we can see that for the first 10 DGMs that assume treatment is not affected by the cluster-level covariate V, all methods have bias much lower than 5%; this is anticipated as throughout analysis models, V is omitted, which shouldn't introduce bias as long as V is not a confounder of the treatment and the outcome. Additionally, the marginal IPTW estimator seems to have a better performance compared to the rest data generating scenarios (where, V is a confounder of the treatment and the outcome), which is also reassuring. The same result applies, e.g., to the first 15 ordered DGMs of figure C.2. A consistent pattern appears across DGMs that do not include V as an independent covariate in the model that generated the treatment values (see figures C.3, C.4, C.5, C.6, C.7, C.8, C.9).

[1] see also: autoplot method for simsum objects developed by Alessandro Gasparini and Ian R. White.

[2] we note that across the current nested loop plots there are a few DGMs missing; however, once updated, we believe that this should not change our overall conclusions.

[3] the black line on the top central part of the graph is printed to illustrate any potential trend of the set parameter values entailed to generate each DGM. However, due to the way these graphs were generated it does not have such an interpretation here - that is because of the current coding system of the DGM via ordered integers; coding via a factorial design for the parameters would allow the (if existent) trend to be illustrated via this line.

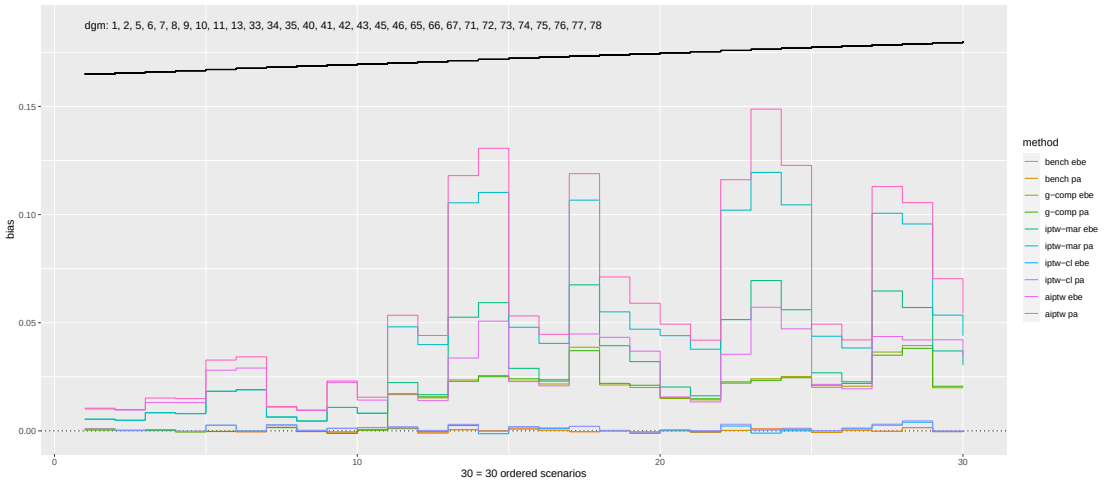


Fig. C.1 Nested loop plot for $p \quad ATE = 0$ and true outcome model with no random effects; simulation of $n_{total} = 1;000$ patients $n_{sim} = 1;000$ times.

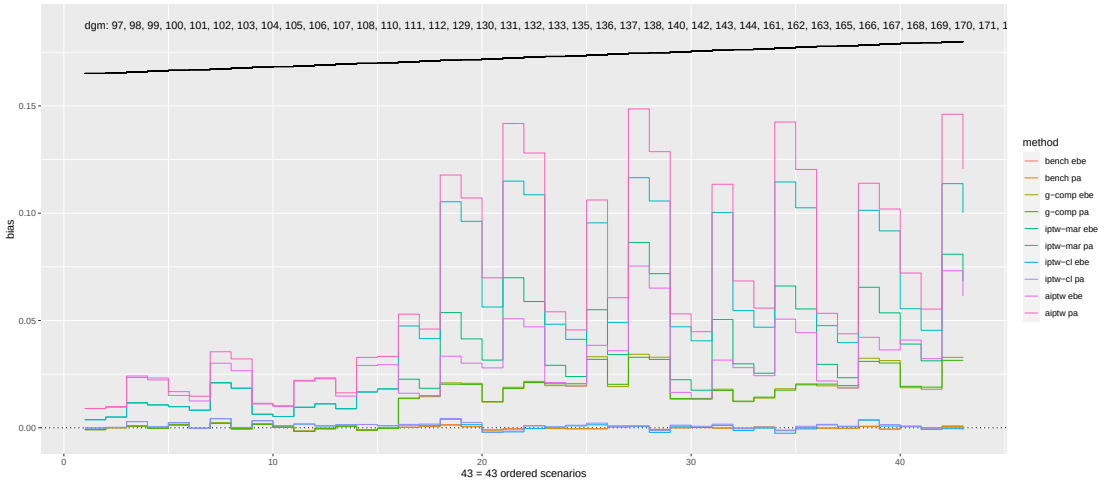


Fig. C.2 Nested loop plot for $p \quad ATE = 0$ and true outcome model with a random intercept; simulation of $n_{total} = 1;000$ patients $n_{sim} = 1;000$ times.

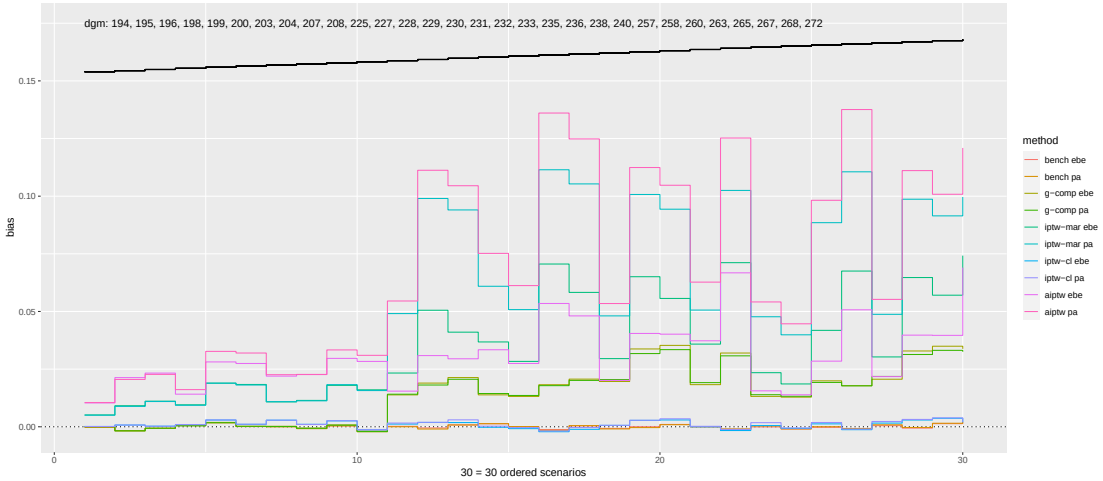


Fig. C.3 Nested loop plot for $p \quad ATE = 0$ and true outcome model with a random intercept and a random slope; simulation of $n_{total} = 1;000$ patients $n_{sim} = 1;000$ times.

And now, for a moderate $p \quad ATE$, and an assumed event rate in the controls of 10%:

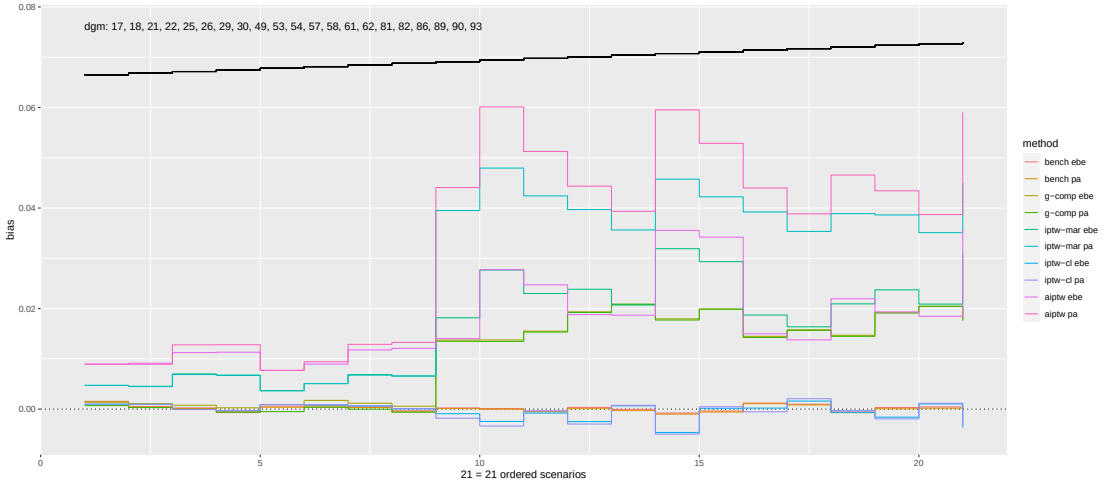


Fig. C.4 Nested loop plot for moderate $p \quad ATE$ (corresponding to $RR = 0{:}8$), event rate in the controls of 10%, and true outcome model with no random effects; simulation of $n_{total} = 1;000$ patients $n_{sim} = 1;000$ times.

And lastly, for a moderate $p \quad ATE$, and an assumed event rate in the controls of 30%:

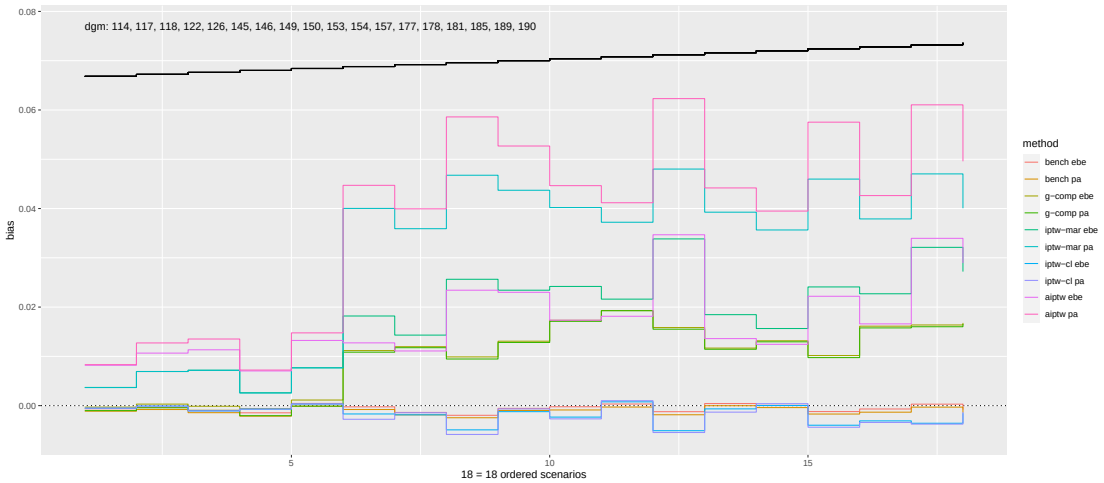


Fig. C.5 Nested loop plot for moderate p ATE (corresponding to RR = 0:8), event rate in the controls of 10%, and true outcome model with a random intercept; simulation of n_{total} = 1;000 patients n_{sim} = 1;000 times.

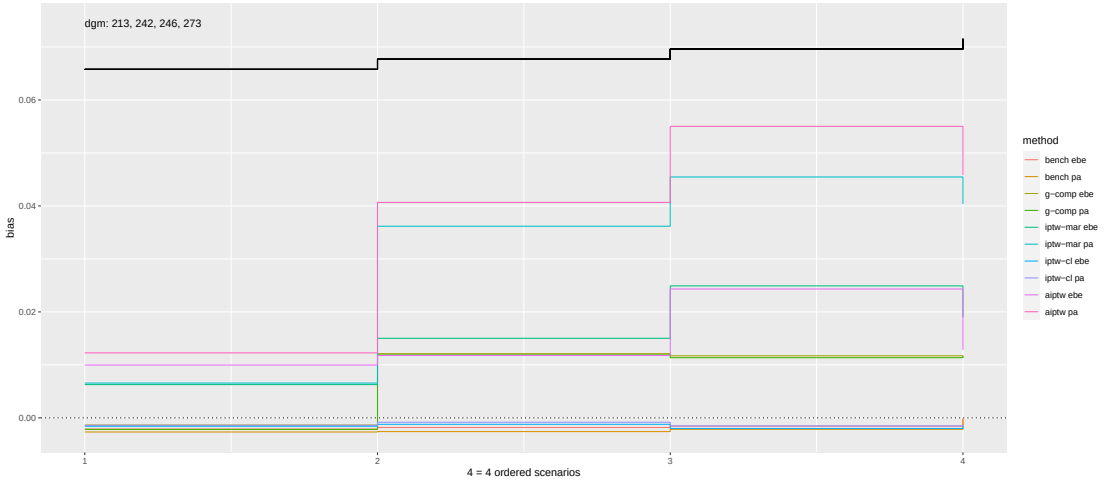


Fig. C.6 Nested loop plot for moderate p ATE (corresponding to RR = 0:8), event rate in the controls of 10%, and true outcome model with random intercept and random slope; simulation of n_{total} = 1;000 patients n_{sim} = 1;000 times.

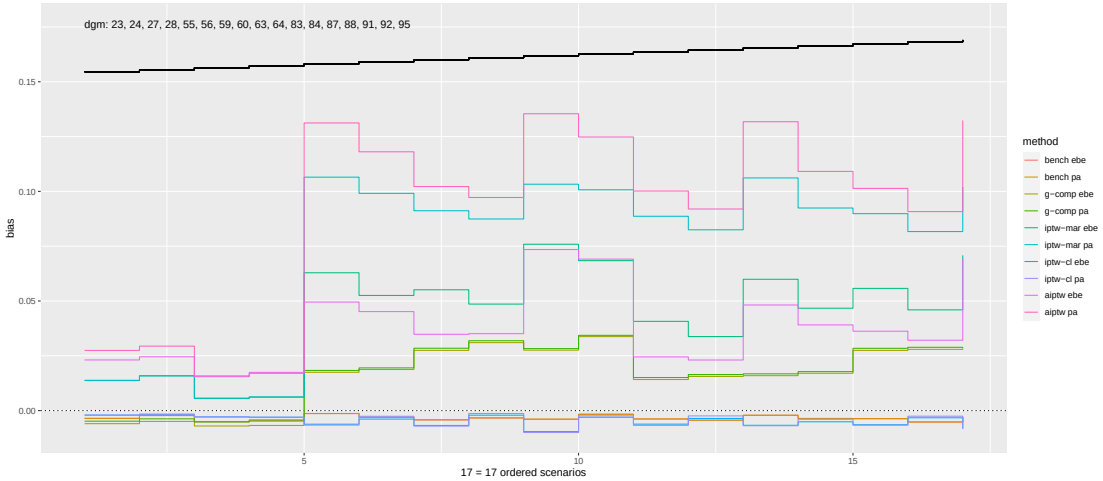


Fig. C.7 Nested loop plot for moderate p ATE (corresponding to RR = 0.8), event rate in the controls of 30%, and true outcome model with no random effects; simulation of n_{total} = 1,000 patients n_{sim} = 1,000 times.

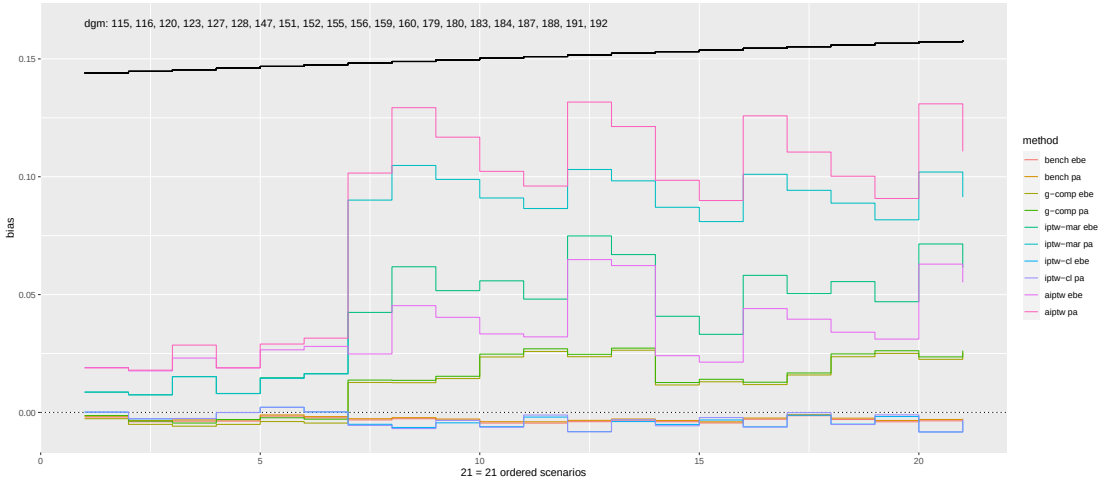


Fig. C.8 Nested loop plot for moderate p ATE (corresponding to RR = 0.8), event rate in the controls of 30%, and true outcome model with a random intercept; simulation of n_{total} = 1,000 patients n_{sim} = 1,000 times.

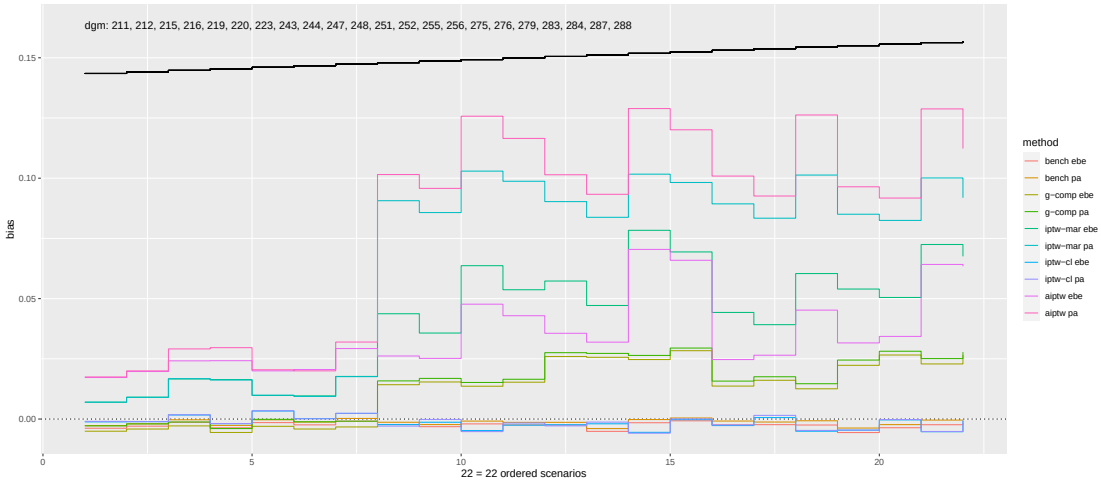


Fig. C.9 Nested loop plot for moderate p ATE (corresponding to $RR = 0{:}8$), event rate in the controls of 30%, and true outcome model with random intercept and random slope; simulation of $n_{total} = 1{;}000$ patients $n_{sim} = 1{;}000$ times.

Appendix D

Additional notes on the causal assumption of consistency

D.1 Consistency: notations and implicit assumptions

According to a different perspective [90], within the consistency assumption, two implicit assumptions are made: i) no variations of treatment level assignment and, ii) no interference. Although the usual practice in the literature is to refer to the combination of the two assumptions under the term stable unit treatment value assumption *SUTVA*, a different take is that, essentially, consistency includes the no interference requirement as well. This could be briefly explained below:

First, recall the most frequently used notation of consistency: $Y_j^{obs} = Y_j(x)$, where, j denotes the individual/patient; x is the level of exposure/treatment; Y_j^{obs} is the observed outcome for individual j; $Y_j(x)$ is the potential outcome for individual j, under exposure/treatment level x. In short, the above assumes that the potential outcome under treatment level x for individual j equals the observed outcome for that individual under the same treatment level. This, however, implicitly assumes that, firstly, in the case of having different ways of assigning a specific treatment level, we would always get the same potential outcome. Let's think of the example of two treatment levels: surgery vs. chemotherapy. Surgery could be received by a different surgeon for different patients. We must therefore assume that the level "surgery" is one and only one, irrespective of the surgeon who conducts the surgery (i.e., no variation of treatment). In the same manner, we can move on to the second implicit assumption made, i.e., no interference. With the former notation, we also assume that the potential outcome of an individual-j under treatment assignment x, will be one and unique, irrespective of anything else (e.g., the treatment assignment of another individual-i). This is often described as "no interference". To be more explicit on the two assumptions of consistency, VanderWeele (2009) suggests a refined notation, like: $Y_j^{obs} = Y_j(x; k_x)$. When $x = X_j$ only for $k_x \in K_x$, And k_x is a means k that the exposure level x can be assigned by; K_x is the different set of means that exposure level x can

be assigned by (e.g., set of different surgeons). Having set the required notation, VanderWeele splits the consistency assumption into the two components we referred to before: $Y_j(x; k_x) = Y_j(x; k_x^0)$ for all $k_x; k_x^0 \in K_x$ C_1 "If condition C_1 holds, then, for any $k_x \in K_x$, we could define $Y_j(x)$ as $Y_j(x) := Y_j(x; k_x)$" Secondly, for each individual-j: For some $k_x \in K_x$, $Y_j^{obs} = Y_j(x; k_x)$ when $x = x_j$ (C2).

Appendix E

Inverse Probability-of-Treatment Weighting (IPTW)

The propensity score is a function of the values of patient characteristics, x_i. PS methods were originally introduced by Rosenbaum and Rubin [73]. However, the so-called IPTW estimator was later proposed by Robins [72]. The IPTW estimator entails two steps:

1. Estimation of the propensity score values
2. Derivation of the IPTW estimator

For the first step, one needs to correctly specify the model for the treatment allocation mechanism, i.e. the PS model. The individual propensity score estimates, i.e., $\hat{e}_i$, can then be generated. In the case of a binary treatment Z ($Z = 1$: treatment, $Z = 0$: no treatment), the usual model of choice for the propensity score is a logistic regression. Once the score is estimated, it can be used to balance characteristics between treatment groups, by re-weighting the individuals by the inverse of the probability of receiving the treatment they actually received. After this procedure, patients are expected to be comparable between the two treatment groups on their observed characteristics.
Then, the IPTW estimators for the marginal mean under treatment, $m_1 = \mathsf{E}[Y_1]$ and under no treatment, $m_0 = \mathsf{E}[Y_0]$ can be derived:

$$\hat{m}_1 = \left(\sum_{i=1}^{n} \frac{Y_i Z_i}{\hat{e}_i} \right) \left(\sum_{i=1}^{n} \frac{Z_i}{\hat{e}_i} \right)^{-1} \tag{E.1}$$

and,

$$\hat{m}_0 = \left(\sum_{i=1}^{n} \frac{Y_i (1 - Z_i)}{(1 - \hat{e}_i)} \right) \left(\sum_{i=1}^{n} \frac{(1 - Z_i)}{(1 - \hat{e}_i)} \right)^{-1} \tag{E.2}$$

The ATE will be:

$$\hat{ATE} = \hat{m}_1 \quad \hat{m}_0 \tag{E.3}$$

These estimators are a modified version of the Horvitz-Thompson estimators originally used in the sampling literature [44] and can be alternatively derived by fitting a weighted regression of the outcome Y on the treatment Z with no other independent variables [1], with the weights being equal to $w_i = 1{=}\hat{e}_i$ for those under treatment and $w_i = 1{=}(1 \quad \hat{e}_i)$ for those under no treatment [43]). Essentially, we weight each individual i by the inverse of their estimated probability of receiving the treatment level they actually received conditional on their characteristics. By doing so, a pseudo-population of individuals is generated. In this pseudo-population, contribution of patients who originally had a high probability of receiving treatment given their characteristics, remains almost the same. On the other hand, those who were underrepresented in the original population, have now their contribution which is up-weighted. In this way, treatment allocation is on average balanced on the measured pre-treatment characteristics X (i.e., in the pseudo-population there is no longer association between the pre-treatment covariates and treatment [43]).

An alternative choice for the weights would be the so-called stabilised weights that include a stabilising term in the numerator which restricts the range of the weights: $sw_i = \frac{\Pr(Z_i)}{\Pr(Z_i|X_i)}$ or more compactly: $sw = \frac{f(Z)}{f(Z|X)}$, where f denotes the probability density function and the denominator is again the estimated values from the PS model. Stabilised weights are expected to have a mean value of 1. Deviations from 1 imply positivity or near positivity violations or model miss-specification. Stabilised weights usually provide narrower 95% confidence intervals than non-stabilised weights. However, the true advantage of stabilised weights is seen when the weighted model (i.e., the model of the outcome fitted on the treatment variable) is non-saturated [2] (e.g., in settings with time-varying or continuous treatments).

Standard errors (SEs) and 95% confidence intervals (CIs) can be generated for the point estimates $\hat{m}_1$ and $\hat{m}_0$ via three different approaches. The first option is to use a robust variance estimator that could be produced by solving the corresponding generalised estimating equations (GEEs) - an automated option widely available in statistical software. However, this choice gives conservative 95% CIs (i.e., they cover the super-population parameter more than 95% of the time), as the robust variance estimator does not consider the uncertainty of the PS estimates (the weights are functions of the estimated PS - *the true PS values are not known*). Another approach is to calculate the variance of the IPTW estimator analytically and a third one could be via the bootstrap method. The last two approaches result in a correct nominal coverage rate and narrower CIs, as they do account for the uncertainty in the PS estimations.

[1] This model is an example of the so-called class of *marginal structural models* (MSM) [72], namely models for the means of the potential outcomes: *marginal*, because they are not conditional on the confounders (models for the population average) and *structural*, because they model *potential* and not observed outcomes.

[2] A conditional model with a number of parameters equal to the number of conditional means in the population is called *saturated*.

Balance assessment

Covariate balance can be checked before weighting using the standardised mean difference (SMD) which is the difference in means between groups, divided by the pooled standard deviation. After weighting, one follows the same procedure, except on weighted means and weighted variances. A rule of thumb is that an absolute value of the SMD that is lower than 0.1 corresponds to sufficient covariate balance. SMDs can be easily checked in a so-called *table one* or in *a plot*. If imbalance is observed, one can redefine the PS model (e.g., add interactions, non-linear terms, etc.) and reassess the balance. SMDs are preferred to statistical tests, as the former do not depend on the sample size. They can be problematic though, for binary variables with very high (or very low) prevalence. For hierarchical (two-level) structures, the same balance diagnostics may be applied. The key distinction is that one may check for balance either across or within clusters, depending on whether the target is the marginal/population average ATE ($p \quad ATE$) or the cluster-specific ATE ($cs \quad ATE$). For the former, balance should be assessed for both individual- and cluster-level characteristics. For the latter, since within cluster, cluster-level characteristics are constant (and therefore, any cluster-level confounding is eliminated), balance needs to be checked only for the individual-level characteristics [75].

Checks for violations of positivity (structural or theoretical *vs* violations in the observed data due to random sampling)

Violations of positivity (i.e., sparsity in the data) can be categorized into two groups: i. structural (or theoretical violations) and ii. positivity violations in the observed data due to random sampling. The former happens when by definition some patients in the population are not allowed to receive a certain treatment (due to clinical characteristics, e.g., susceptibility to the particular treatment). The latter occurs when certain subgroups of patients rarely or never receive the treatment in the observed data at hand due to random sampling. The last type can improve with an increase in the sample. In general, positivity violations may lead to an increase in bias accompanied with or without a variance inflation [65].

The main way to identify non-positivity is to bear in mind the clinical question and the related reasons that could potentially lead to such a violation. To investigate for indications of positivity violations empirically, one should examine the distribution of the propensity score and look for extreme values in the weights. However, the absence of extreme weight values does not necessarily ensure that positivity holds [21].

In the case of large weight values (indicative of near positivity violations), one option is trimming the tails (i.e., removing subjects who have extreme values of the PS [3]). But this tactic usually changes the population. Another option would be truncating the large weights (i.e., determining a maximum

[3] A common trimming tactic: i. Remove treated subjects whose propensity scores are above the 98^{th} percentile from the distribution among controls and ii. Remove control subjects whose propensity scores are below the 2^{nd} percentile from the distribution of treated subjects

allowable weight value (e.g., 100) or a percentile (e.g., 99^{th}) - and if the weight is larger than the maximum allowable, set it to the maximum allowable value, etc.). This procedure introduces some bias, but reduces the variance. A few other approaches would be: restriction of the covariate adjustment set (although, this could potentially lead to confounding bias) or re-definition of the causal effect of focus - this time, as an effect of treatments that do not result in positivity violations, mainly for the case of structural non-positivity (although, this is a compromise between the original target causal effect and the extent to which it can be identifiable).

Appendix F

Standardisation and the parametric G-formula

Unless otherwise stated, this section was based on Chapter 13 of [43].

Standardisation is another method to eliminate confounding in observational studies. Under the required (*identifiability*) assumptions, one can estimate the causal effect of some treatment on some outcome from observational data. It can be mathematically proven that standardisation and IPTW are equivalent and give identical results when it comes to the non-parametric approach [1]. However they provide similar but not identical results when a parametric approach is used, because the former models the outcome mechanism whereas the latter models the treatment allocation distribution. Nonetheless, the validity of the inferences from both methods relies on the same identifiability assumptions seen in section 2.1.3.

Assume an observational cohort study with a binary treatment of interest and the other quantities defined as throughout the document. Recall that our aim is to estimate the ATE (*or the ATT*) from our observed data. Hence, the ATE requires the estimation of $m_1 = \mathsf{E}[Y_1]$ and $m_0 = \mathsf{E}[Y_0]$. The expected value of our outcome if the whole population were to receive treatment, or in other words m_1, can be estimated from the so-called standardised mean, namely:

$$\mathsf{E}[Y_1] = \sum_x \mathsf{E}[Y|Z = 1; X = x] \quad \Pr[X = x] \tag{F.1}$$

Equation F.1 corresponds to the case where the potential confounders X are assumed to be discrete. For continuous variables X we can simply modify formula F.1 by substituting $Pr[X = x]$ with the probability density function (PDF) $f_X[x]$ and the sum by an integral. Additionally, the standardised mean in the treated illustrated by F.1 is for the specific case scenario, where there is no censoring. In

[1] For a proof, consult Technical Point 2.3, p 24 in [43].

the presence of censoring, the first term of F.1 (conditional mean outcome) is done in the uncensored treated. If we denote the censoring variable with C, then:

$$\mathsf{E}[Y_1] = \sum_x \mathsf{E}[Y|Z = 1; C = 0; X = x] \quad Pr[X = x] \tag{F.2}$$

The standardised mean in the treated (Z=1) is defined as the weighted average of the conditional mean outcomes $\mathsf{E}[Y|Z = 1; X = x]$ in each different stratum of x of the potential confounders X using as weights the probability of having the individual characteristics x in the study population, namely, $Pr[X = x]$ (eq F.1). To estimate μ_1 via the equation F.1 we need to estimate the two aforementioned quantities, namely the conditional means and the weights $Pr[X = x]$. The estimation of the probabilities $Pr[X = x]$ would be possible in the case of one confounder with a finite, and, often small, number of levels. However, the non-parametric approach is challenging or even impossible in the presence of numerous confounders, often with multiple levels (i.e., high-dimensional data). In that case one should resort to modelling instead. The parametric version of standardisation could be divided into 4 steps:

> [i.]Model the outcome as a function of the treatment and the confounders. Predict the outcome in the whole population as if everyone was treated (respectively untreated). Average these predictions for the treated (respectively for the untreated). Contrast these averages according to the estimand of interest (e.g. risk difference).

The first step is to model the outcome Y against the set of measured pre-treatment baseline characteristics X and the treatment. It is highlighted once more that the aim of this step is to correctly model the outcome Y as a function of the measured pre-treatment characteristics X and treatment of interest Z. By fitting for instance a linear regression for continuous outcomes or a generalised linear model (e.g., for binary or ordinal ones), or conducting time-to-event analysis, we can model the below quantity of interest:

$$\mathsf{E}[Y|Z = z; X = x] \tag{F.3}$$

Then (step ii) one could predict the individual outcome values for each different combination of values x if our whole dataset was assigned to the treated and if our whole dataset was assigned to the untreated respectively. Step iii. would be standardisation by averaging across each dataset, or, in other words calculation of the sample averages of the predicted values correspondingly for the treated (Z=1) and the untreated (Z=0). Then the difference in the standardised means would provide the ATE. For the calculation of the SEs for the 95% confidence intervals, bootstrapping would be an option.

Appendix G

Doubly robust methods

G.0.1 Inverse probability-of-treatment weighting: regression adjustment (IPTW-RA)

IPTW-RA combines outcome regression with treatment modelling. It consists of the following steps: Firstly, one models the treatment allocation mechanism against the potential confounders (i.e., PS model) and derives the weights (IPTW). Secondly, one models the outcome against the postulated confounders and treatment, using the IPTWs derived from the first step as weights. Thirdly, one predicts the expected response if the whole population were to be treated and untreated, respectively. Lastly, one follows the standardisation step, such as in regular standardisation. In this way, one may obtain the estimates for m_1 and m_0, and correspondingly for the difference between them. SEs and CIs are generated via the bootstrap [84].
IPTW - RA provides consistent estimates if either i. the outcome regression function or ii. the PS model is correctly specified [47, 85, 79]. When the former is not correct, the latter corrects for it, and vice versa. However, it is the least robust in comparison to the other DR methods.

G.0.2 Augmented inverse probability-of-treatment weighting (AIPTW)

AIPTW is a DR estimator originally introduced by Robins and colleagues [71]. In general, IPTW may be variable in the case of extreme values of the weights. AIPTW is more robust to model mis-specification and less variable than the IPTW. The AIPTW estimator is constructed by including the IPTW and standardisation model into one. The intuition of the AIPTW estimator for the estimation of $m_1 = \mathsf{E}[Y(1)]$ is briefly presented below. The estimation of $m_0 = \mathsf{E}[Y(0)]$ would be quite similar, and, afterwards the difference in means due to treatment would provide the ATE. Let us suppose an observational study of n individuals, an outcome Y, a binary treatment Z and some measured predictors X. Based on the idea to combine a propensity score estimator, such as IPTW from 2.15 and a maximum likelihood estimator from a regression model of the outcome, fitted for the estimation of the standardised mean in F.1, the AIPTW estimator for the estimation of m_1 can be derived [46]:

$$\hat{m}_{1;AIPTW} = n^{-1} \sum_{i=1}^{n} \left(\frac{Z_i Y_i}{e(X_i;\hat{a})} - \frac{(Z_i - e(X_i;\hat{a}))}{e(X_i;\hat{a})} m_1(X_i;\hat{b}_1) \right) \tag{G.1}$$

Where $e(X_i;\hat{a})$ is the estimated individual PS values expressed as a function of the measured predictors and the estimations for the parameters a of the PS model, $\hat{a}$ often obtained by a logistic regression. Also, $m_1(X_i;b_1)$ would be the assumed regression model for the true relationship between the vector of covariates and the outcome within the treated. Hence, the vector of $\hat{b}_1$ is the vector of parameter estimates for the outcome regression model. The previous equation can be approximated by:

$$\mathrm{E}[Y(1)] + \mathrm{E}\underbrace{\left[\frac{\{Z - e(X;a)\}}{e(X;a)} \{Y(1) - m_1(X;b_1)\} \right]}_{augmentation} \tag{G.2}$$

When n is large, the sample average in G.1 estimates the population average (G.2). The first term in equation G.2, $\mathrm{E}[Y(1)]$, is the average outcome in the treated. The second term is an augmentation term. If the augmentation term reduces to zero, then equation G.2 will estimate m_1 (i.e., the expected response if the whole population were to receive treatment). A more intuitive explanation for the derivation of the estimator in G.1 could be: AIPTW includes two quantities - the first, is the quantity of interest, i.e., the mean outcome if the whole population were to be treated. The second quantity is essentially the product of two bias terms - the bias of the PS model and the bias of the outcome regression model. If the PS model is correct, then bias from this model will be zero, and it will "zero out" the bias from the outcome regression model. Likewise, if the outcome regression model is correctly specified (i.e., zero bias), it will "zero out" the bias from the PS model. In other words, if either the PS or the outcome regression model is correctly defined, the estimator from G.1 will give an unbiased estimate - the so-called *doubly robust property*. A very comprehensive proof of this property can be found in the Web Appendix of the work of Jonsson Funk and colleagues [46], while exceptional but more technical references would be the works of Bang and Robins [12], Tsiatis [84] or Van der Laan and Robins [87].

Following the same logic, the AIPTW for the estimation of m_0 will be:

$$\hat{m}_{0;AIPTW} = n^{-1} \sum_{i=1}^{n} \left(\frac{(1 - Z_i)Y_i}{1 - e(X_i;\hat{a})} + \frac{\{Z_i - e(X_i;\hat{a})\}}{1 - e(X_i;\hat{a})} m_0(X_i;\hat{b}_0) \right) \tag{G.3}$$

And, to estimate the ATE:

$$\hat{ATE} = \hat{m}_{1;AIPTW} - \hat{m}_{0;AIPTW}$$

Lastly, we present the corresponding formulas for each individual for the expected outcome under treated and untreated conditions derived from eq. G.3, G.1 below:

Abbreviations: AIPTW, Augmented Inverse Probability-of-Treatment Weighting; $\hat{e} = P(Z = 1|X)$; Z= treatment; $Y_{Z=0}$ and $Y_{Z=1}$ observed outcome among individuals with $Z = 0$ and $Z = 1$, respectively; $\hat{Y}_0 = \mathrm{E}(Y|Z = 0;X)$ = predicted outcome given $Z = 0$; $\hat{Y}_1 = \mathrm{E}(Y|Z = 1;X)$ = predicted outcome

Table G.1 Equations for the expected response under treated ($AIPTW_1$) and untreated ($AIPTW_0$) conditions for each individual in the population

	$AIPTW_1$	$AIPTW_0$
General form	$\frac{Y_{Z=1} \quad Z}{\hat{e}} \quad \frac{\hat{Y}_1(Z \quad \hat{e})}{\hat{e}}$	$\frac{Y_{Z=0} \quad (1 \quad Z)}{1 \quad \hat{e}} + \frac{\hat{Y}_0(Z \quad \hat{e})}{1 \quad \hat{e}}$
Among $Z = 1$	$\frac{Y_{Z=1}}{\hat{e}} \quad \frac{\hat{Y}_1(1 \quad \hat{e})}{\hat{e}}$	$\hat{Y}_0$
Among $Z = 0$	$\hat{Y}_1$	$\frac{Y_{Z=0}}{1 \quad \hat{e}} \quad \frac{\hat{Y}_0\hat{e}}{1 \quad \hat{e}}$

given $Z = 1$; individual subscript i is suppressed for readability; table inspired by Funk et al., 2011 [46].

G.0.3 Targeted maximum likelihood estimation (TMLE)

This subsection was based on [58, 88].

TMLE is an estimation technique that offers an optimal exchange between variance and bias in the estimation of the target parameter (e.g., ATE). TMLEs belong to the class of DR and efficient estimators. The core idea is that for the ATE, TMLE takes initial estimates of $\mathsf{E}[YjZ;X]$ and $Pr(ZjX)$ and afterwards incorporates a substitution *targeting* step which optimises the bias-variance trade-off for the parameter of focus (the ATE in our case). Statistical inference and 95% CIs are based on the efficient influence curve (EIC) theory, but at this stage we will not delve into more technical details. Briefly, the first step of the algorithm is calculating the so-called Q-model (i.e., outcome regression model as a function of treatment and potential confounders), which is essentially the application of standardisation, to predict the outcome if the whole population were treated and untreated respectively. This gives an initial estimate $\bar{Q}^0(Z = 1;X)$ and $\bar{Q}^0(Z = 0;X)$, and of the difference between the two, i.e., the ATE. Then, estimation of the PS follows. To counteract the potential residual bias from the association of treatment and potential confounders in our initial estimate $\bar{Q}^0$ from the Q-model, we create the so-called *clever covariates* and estimate the *fluctuation parameter* $\mathsf{e} = (\mathsf{e}_0;\mathsf{e}_1)$. Clever covariates are similar to the traditional inverse probability-of-treatment weights. A regression model of the outcome Y against the logit of the initial prediction $\bar{Q}^0$ as an offset and the clever covariates (denoted usually as $H(1;X)$, $H(0;X)$) as independent variables is fitted. And this is the key part: if there is actually residual confounding (relationship between the treatment Z and the potential confounders) that was not captured by the initial model, then the clever covariates will account for that by providing an updated estimate (if the PS model is correctly specified). If the initial model was correctly specified, then the fluctuation parameter estimate will be close to zero, because in that case, the PS model would not provide additional information to the initial estimate. There is also the case where the fluctuation parameter estimate might end up close to zero (because the PS might not give additional information to the initial estimate), given although that the initial Q-model is misspecified.

SEs may be calculated via the estimation of the variance of the so-called EIC estimate for the ATE based on asymptotic theory.
The TMLE algorithm may have more than one updating steps in more complex settings. Because of the clever covariates inclusion, TMLE tends to perform better than other DR approaches. In addition to that, TMLE can incorporate machine learning techniques, making it even more robust to model mis-specification. These features, make this method a lot promising when it comes to extra layers of complexity (e.g., hierarchical structures, [11]).