# Methodological obstacles in causal inference: confounding, missing data, and measurement error

Penning de Vries, B.B.L.

Methodological obstacles in causal inference

Bas Penning de Vries

# Methodological obstacles in causal inference: confounding, missing data, and measurement error

Bas Penning de Vries

Methodological obstacles in causal inference: confounding missing data,
and measurement error

PhD thesis, Leiden University, the Netherlands

Printed by UFB, Leiden University

# Methodological obstacles in causal inference: confounding, missing data, and measurement error

**Proefschrift**

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 25 januari 2022
klokke 15.00 uur

door

**Bas Bernd Lodewijk Penning de Vries**

geboren te Hilversum
in 1992

## Promotor

Prof.dr. R.H.H. Groenwold

## Copromotor

Dr. M. van Smeden (University Medical Center Utrecht)

## Leden promotiecommissie

Prof.dr. S. le Cessie
Prof.dr. E.W. Steyerberg
Prof.dr. K.C.B. Roes (Radboud University Medical Center)
Prof.dr. O.H. Klungel (Utrecht University)

# Table of contents

# 1

## General introduction and outline of thesis

What it means for something to cause something else has long been, and still is, a topic of debate among philosophers. In epidemiology and other sciences, a now-dominant approach to causality—the *counterfactual* or *potential outcomes framework*—considers causal claims that address what-if questions, claims about the consequences of hypothetical—possibly contrary-to-fact—actions (Neyman et al., 1935; Rubin, 1974; Holland, 1986, 1988; Pearl, 2009; Hernán and Robins, 2020). Statements of this form arise naturally in medicine, for example, where we are often faced with the problem of having to choose between treatment options for a patient and where—to guide decision making—we might ask what would happen with the patient's well-being if we choose one treatment option versus another.

### 1.1 Causal inference

Within the counterfactual outcomes framework, a *causal effect* is defined in terms of a contrast between at least two hypothetical actions. That at most one of mutually exclusive actions will actually become factual means that causal effects are not directly apparent. To proceed in our endeavour to quantify a causal effect, we start by looking for a way to equate our target quantity, the *estimand*, with a known function of the distribution of factual, real-world variables (Petersen and Van der Laan, 2014; Ahern, 2018). If we succeed in finding such a way, the estimand is said to be *identifiable* (from this distribution of factuals) and the expression that describes how it relates to the distribution is an *identifiability expression*. If the distribution of factuals is compatible with at least two values of the estimand, the estimand is said to be non-identifiable and no identifiability expression exists.

The identifiability expression forms the basis for estimation. Where identification connects the estimand with the theoretical (sometimes called 'population') distribution of factual variables, estimation concerns how these theoretical distributions in turn relate to empirical distributions or finite random samples. An *estimator* is a function of a random sample that is intended to offer a close approximation of the estimand. Because the estimator is a function of a random sample, its output (sometimes also labeled the estimator) is itself a random variable. Given a fixed realisation of the sample, the output of an estimator is fixed and known as an *estimate*.

## 1.2 Obstacles in causal inference

There are a number of obstacles in providing 'good' estimates of causal effects. Some of these are purely statistical and relate only to estimation but others complicate identification as well as estimation. Three of these—confounding, missing data, and measurement error—are the main themes of this thesis and are described briefly below by way of an example.

Suppose that the interest lies with a contrast between someone's risk of sustaining a health-related event had this person been subjected, possibly contrary to fact, to exposure $A = 1$ (e.g., initiating a certain drug treatment, immediately after diagnosis with some condition) and the risk of the possibly contrary-to-fact situation where $A$ was set to 0 (e.g., not initiating the drug treatment immediately after diagnosis). To clarify which hypothetical or counterfactual situation we are referring to, we write $Y(0)$ for the indicator of the event of interest that would be realised had the person been exposed to $A = 0$ and we likewise denote by $Y(1)$ the event indicator that would be realised had the person been exposed to $A = 1$. In the literature, and also in this thesis, rather than enclosing it within parentheses, a reference to a counterfactual situation is sometimes written in superscript (e.g., $Y^0$) or subscript (e.g., $Y_0$) to indicate the corresponding counterfactual version of a variable. In this example, the event indicators each take one of two levels, 1 in case the event of interest takes place and 0 otherwise. The interest can now be succinctly written as

$$\Pr(Y(1) = 1) \text{ versus } \Pr(Y(0) = 1). \tag{1.1}$$

### *1.2.1 Confounding*

The simplest attempt at identifying the components of the contrast (1.1) is to replace the counterfactual event risk $\Pr(Y(a) = 1)$ with $\Pr(Y = 1|A = a)$ for $a = 0, 1$, yielding

$$\Pr(Y = 1|A = 1) \text{ versus } \Pr(Y = 1|A = 0). \tag{1.2}$$

The event indicator $Y$ is the factual (observed), real-world outcome variable. But under what conditions are (1.1) and (1.2) actually equivalent, or when is $\Pr(Y(a) = 1)$ identified by $\Pr(Y = 1|A = a)$? Three key identifiability conditions are consistency, exchangeability, and positivity.

*Consistency* connects the counterfactual variables of interest with the factual, real-world variables (Cole and Frangakis, 2009; VanderWeele, 2009; Pearl, 2010). It means that the counterfactual version of an outcome variable (e.g., characterising the well-being of a patient) under a given hypothetical action coincides with its factual version if this hypothetical action agrees with (our impression of) the real world. For example, if a patient is known to have received treatment with a particular drug, consistency implies that the patient's well-being is the same as the patient's well-being that would have been realised had the patient been assigned this treatment. This seemingly trivial condition has two noteworthy subtleties. First, a prerequisite is that the hypothetical actions of interest are sufficiently well-defined. There may be many variations on "assigning drug treatment" (e.g., in the dosing or timing of the treatment) and their impact on the patient's well-being need not be the same. The actions are sufficiently well-defined if there is no ambiguity about the variation or all possible variations equally affect the outcome variables of interest (i.e., there is treatment variation irrelevance). Second, it may be that a patient's treatment is misclassified, resulting in a wrong impression about the real world. In turn, the patient's well-being need not coincide with its counterfactual counterpart that would be realised had the received the registered treatment. We will reconsider misclassification below. For now, let us assume that there is no such misclassification and that $Y(a) = Y$ if $A = a$, for $a = 0, 1$, so that

$$\begin{aligned}
\Pr(Y(a) = 1) &= \Pr(Y(a) = 1|A = a)\Pr(A = a) \\
&\qquad + \Pr(Y(a) = 1|A = 1 - a)\Pr(A = 1 - a) \\
&\qquad\qquad\qquad\qquad \text{(by the law of total probability)} \\
&= \Pr(Y = 1|A = a)\Pr(A = a) \\
&\qquad + \Pr(Y(a) = 1|A = 1 - a)\Pr(A = 1 - a). \quad \text{(by consistency)}
\end{aligned}$$

The application of the law of total probability above and (1.2) require that the conditional probabilities given $A = 1$ or $A = 0$ are defined, i.e., that $A = 1$ and $A = 0$ have positive probability (*positivity*).

Note that the right-hand side of the above equality has only one term that contains a counterfactual outcome. Under exchangeability, this term can be replaced and the above expression can be turned into an identifiability expression. *Exchangeability*, or 'no confounding', here means that $Y(a) \perp\!\!\!\perp A$—shorthand for $Y(a)$ is independent of $A$—for $a = 0, 1$, so that

$$\begin{aligned} \Pr(Y(a) = 1|A = 1 - a) &= \Pr(Y(a) = 1|A = a) && \text{(by exchangeability)} \\ &= \Pr(Y = 1|A = a). && \text{(by consistency)} \end{aligned}$$

Hence, under consistency, exchangeability and positivity,

$$\begin{aligned} \Pr(Y(a) = 1) &= \Pr(Y = 1|A = a)\Pr(A = a) \\ &\quad + \Pr(Y = 1|A = a)\Pr(A = 1 - a) \\ &= \Pr(Y = 1|A = a)[\Pr(A = a) + \Pr(A = 1 - a)] \\ &= \Pr(Y = 1|A = a) \end{aligned}$$

and, so, (1.1) and (1.2) are equivalent. When the exchangeability condition is violated, however, $\Pr(Y(a) = 1|A = a) \neq \Pr(Y(a) = 1|A = 1 - a)$ for $a = 0$ or $a = 1$, and therefore $\Pr(Y = 1|A = 0)$ and $\Pr(Y = 1|A = 1)$ might not equal $\Pr(Y(0) = 1)$ and $\Pr(Y(1) = 1)$, respectively.

Departure from identification of an estimand, brought about for example by a violation of the exchangeability condition, may have a knock-on effect on the properties of an estimator. Given a sample of $n$ exposure-outcome pairs $(A_i, Y_i)$, natural estimators of the components of the contrast (1.2) are

$$\sum_{i=1}^{n} \frac{A_i}{\sum_{j=1}^{n} A_j} Y_i \text{ and } \sum_{i=1}^{n} \frac{1 - A_i}{\sum_{j=1}^{n}(1 - A_j)} Y_i. \tag{1.3}$$

Suppose that the number of exposures is fixed at $m$ (i.e., $\sum_{i=1}^{n} A_i = m$). If for $a = 0, 1$, $Y_i|A_i = a$ has the same distribution as $Y|A = a$, then

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{n} \frac{A_i}{\sum_{j=1}^{n} A_j} Y_i\right] &= \frac{1}{m}\sum_{i=1}^{n} \mathbb{E}[A_i Y_i] \\ &= \frac{1}{m}\sum_{i=1}^{n} \mathbb{E}[Y_i|A_i = 1]\Pr(A_i = 1) \\ &\qquad \text{(by the law of total (or iterated) expectation)} \end{aligned}$$

$$
\begin{aligned}
&= \Pr(Y=1|A=1)\frac{1}{m}\sum_{i=1}^{n}\Pr(A_i=1)\\
&= \Pr(Y=1|A=1)\frac{1}{m}\mathbb{E}\left[\sum_{i=1}^{n}A_i\right]\\
&= \Pr(Y=1|A=1).
\end{aligned}
$$

We therefore say that the estimators of (1.3) are unbiased for the components of (1.2) if $Y_i|A_i \sim Y|A$ and in turn for the components of (1.1) if additionally the identifiability assumptions (consistency, exchangeability and positivity) are met. More generally, given fixed $\theta$, we say that an estimator $\hat{\theta}$ is unbiased for a quantity $\theta$ if $\mathbb{E}[\hat{\theta}] = \theta$, and we refer to $\mathbb{E}[\hat{\theta} - \theta]$ as the bias of the estimator $\hat{\theta}$ (relative to $\theta$). This notion of bias relates to estimation. However, in this thesis we also occasionally use the term bias to refer to the distance between the quantity that is identified by the identification strategy and the value of the estimand. Of note, there is often a connection between the bias of an estimator and the bias of an identification strategy: in many cases, the former converges in some sense to the latter as the sample size on which the estimator is based increases. Confounding is often labeled a 'source of bias'—an apt description regardless of the notion of bias.

Bias, or the lack thereof, is only one of many properties of an estimator that describe how 'good' an estimator approximates a target quantity. The other properties include, but are not limited to, variance and mean squared error, which are addressed in other chapters of this thesis.

### *1.2.2 Missing data*

Identification is a relative notion—it is relative to a set of factual variables whose distribution we seek to connect to the estimand. Including in this set of factuals variables that are not observed by the end of data collection has little practical value and, so, we eventually restrict our attention to the observed part. Before we do, however, it is often useful—although not always necessary—to consider, as an intermediate step, identification from a set of variables that may not be fully observed. If the distribution of these variables can be inferred (or 'recovered') from the distribution of the observed part, then identifiability from the former implies identifiability from the latter.

This insight motivates the classification of missingness as either *missingness* that is *completely at random* (MCAR), *at random* (MAR), or *not at random* (MNAR) (Rubin, 1976). Consider a sequence $(X_1, ..., X_p)$ of $p$ variables and an equally long sequence $(R_1, ..., R_p)$ of response indicators, with $R_i = 0$ if $X_i$ is observed and $R_i = 1$ otherwise, $i = 1, ..., p$. Data are said to be MAR relative to a realisation $(r_1, ..., r_p)$ of $(R_1, ..., R_p)$ and realisations $(x_i : r_i = 1)$ of $(X_i : r_i = 1)$, if for all levels $(x_i : r_i = 0)$ of $(X_i : r_i = 0)$,

$$\begin{aligned}&\Pr(R_1 = r_1, ..., R_p = r_p | X_1 = x_1, ..., X_p = x_p)\\ &\quad = \Pr(R_1 = r_1, ..., R_p = r_p | X_i = x_i : r_i = 1);\end{aligned}$$

and MCAR relative to this realisation (a stronger condition) if

$$\Pr(R_1 = r_1, ..., R_p = r_p | X_1 = x_1, ..., X_p = x_p) = \Pr(R_1 = r_1, ..., R_p = r_p).$$

Missingness that is not 'at random' (or 'completely at random') is 'not at random'. An interesting special case where MAR is always satisfied is the case where the first $j$ variables, $1 \leq j \leq p$, are always observed (i.e., $R_1 = ... = R_j = 1$) and the missingness of the other $p - j$ variables satisfies

$$(R_i : j < i \leq p) \perp\!\!\!\perp (X_i : j < i \leq p) | (X_i : 1 \leq i \leq j).$$

For this special MAR condition, it is easy to determine whether the joint distribution of $(X_1, ..., X_p, R_1, ..., R_p)$ is recoverable from the distribution of its observed part: the distribution of the partially observed (i.e., last $p - j$) variables given the always observed (i.e., first $j$) variables is obtained by conditioning on $R_1 = ... = R_p = 1$.

It is interesting to consider counterfactual outcomes that are strictly 'non-factual' as missing variables. If $Y(0) = Y$ whenever $A = 0$ and there are no missing factuals, the missingness of $Y(0)$ in (1.1) is fully determined by $A$: $Y(0)$ is observed if and only if $A = 0$. The missingness is therefore 'at random'. The counterfactual outcome probability $\Pr(Y(0) = 1 | A = 1)$ is however not identified by $\Pr(Y(0) = 1 | A = 1, R_0 = 1)$, with $R_0$ denoting the response indicator for $Y(0)$, because $R_0 = 1$ if and only if $A = 0$ and, since $\Pr(A = 1, A = 0) = 0$, $\Pr(Y(0) = 1 | A = 1, A = 0)$ is not defined.

### *1.2.3 Measurement error*

Measurement error arises when our impression of a variable's value is different from its actual value. The variable can be considered unobserved and its potentially wrong impression forms another, observed variable.

A special form of measurement error is misclassification, which means that our measurement or impression of a categorical variable is inaccurate. As for the above example, exposure misclassification means that our measurement or 'impression' $A^*$ of $A$ does not always coincide with $A$ itself. While $Y(a)$ might equal $Y$ if $A = a$ for $a = 0, 1$, it is possible that $Y(a) \neq Y = Y(A)$ if $A^* = a$. This is sometimes described as a possible violation of the consistency assumption (Gravel and Platt, 2018); the counterfactual outcome $Y(a)$ need not coincide with the observed $Y$ even if $A^* = a$. Upon replacing $A$ with $A^*$, (1.2) becomes

$$\Pr(Y = 1|A^* = 1) \text{ versus } \Pr(Y = 1|A^* = 0). \tag{1.4}$$

A simple yet naïve approach is to take (1.4) as the basis for inference about (1.1). However, that (1.1) and (1.4) are equivalent is not evident and may not be true. More generally, like confounding and missing data, measurement error is an obstacle in causal inference.

## 1.3 Objective and outline of thesis

There are many approaches to handling the obstacles of confounding, missing data and measurement error. To address confounding, these approaches include traditional regression analyses and the more modern propensity score methods such as propensity score matching and inverse probability weighting (Rosenbaum and Rubin, 1983; Robins et al., 2000). These methods rely on the availability of other variables, covariates, such that conditional on these covariates, there is exchangeability. Other methods, like instrumental variable analysis and negative control methods, rest on different assumptions (Greenland, 2000; Lipsitch et al., 2010). For missing data, simply discarding incomplete records has long been the default approach. More principled approaches include expectation-maximisation, multiple imputation, and inverse probability weighting (Dempster et al., 1977; Rubin, 1987; Robins et al., 2000). Lastly, the impact of measurement error may be mitigated by regression calibration, simulation extrapolation (SIMEX), latent variable modelling, or inverse probability weighting (Buonaccorsi, 2010; Gravel and Platt, 2018).

The development and study of the properties of methods to overcome the abovementioned methodological obstacles in epidemiology is an active area of research with many open questions. The aim of this thesis is to contribute to this research and to provide more insight into the properties of methods for confounding, missing data and measurement error.

The outline for the remainder of this thesis is as follows. We start by considering in **chapter 2** the task of reviewing the existing literature to gauge the current state of knowledge about existing methodologies, identify gaps or to provide a starting point for guidance development. The chapter gives an appraisal of methods for gathering information about the use of methods for the study of causal effects. In all subsequent chapters, we zoom in on methods of the latter kind. In **chapters 3 and 4**, we address the concern that the combination of multiple imputation for missing data and propensity score methods for confounding has worse performance than might be expected from how they perform in isolation. In doing so, we compare two strategies from epidemiological practice for implementing propensity score methods to multiply imputed data sets and we give guidance on which is to be preferred. The study of propensity score methods and missing data is continued in **chapter 5**, which focuses on a class of machine learning methods, classification and regression trees (CART), for estimating propensity scores in the presence of missing covariate data. **Chapter 6** then turns to propensity score matching and missing outcome data. The chapter illustrates that when baseline exchangeability is achieved through propensity score matching, bias might result from restricting downstream analysis to the subset of individuals who have not dropped out of the study by the administrative study end. In **chapter 7**, we consider missing data mechanisms that are governed by study design. In studies on the effects of time-varying exposures, adequate information on time-varying participant characteristics might help mitigate time-dependent confounding. However, the frequency with which these characteristics are measured may be inadequate and participant characteristics are sometimes (wrongly) assumed to remain constant in periods of no measurement (i.e., there is measurement error). The chapter illustrates the impact of design choices regarding data collection. Measurement error is also a dominant theme in **chapter 8**, in which a weighting method for simultaneous adjustment for confounding and joint exposure-outcome misclassification is developed. The method relies on standard identifiability assumptions, such as exchangeability within levels of a collection of partially observed variables, consistency, positivity, and MAR. In many studies on causal effects, however, there are often concerns that standard identifiability assumptions are violated. Negative controls are a tool with the potential to detect or correct for confounding that is explained by fully unobserved variables. This is the topic of **chapter 9**. The counterfactual outcomes framework and attempts to identify estimands have become increasingly popular in much of the causal inference literature, including the literature on negative controls. Case-control studies have not yet enjoyed this trend. In **chapter 10**, we reconsider this family of designs and recast classical concepts, assumptions and principles from a modern perspective. It is shown how and when a variety of causal estimands can be identified with these study designs. Causal inference and prediction are two areas of epidemiology that are increasingly seen as overlapping. In precision medicine, it is not uncommon for treatment assignment decisions to be based on 'prognostic scores', predictions of the outcome of interest that would be realised if the treatment were withheld. **Chapter 11** deals with this topic and emphasises that in order to obtain optimal results, the counterfactual outcomes under both treatment levels, 'treatment' and 'no treatment', should be considered. The methodological obstacles that we encounter in causal inference, including confounding, missing data and measurement error, are therefore relevant in that context too. To conclude, in **chapter 12**, we present a summary of the previous chapters, along with a general discussion of this thesis in the light of the existing literature, with suggestions for future research.

## 1.4 References

Ahern, J. (2018): "Start with the "C-word," follow the roadmap for causal inference," *American Journal of Public Health*, 108, 621.

Buonaccorsi, J. P. (2010): *Measurement error: models, methods, and applications*, Boca Raton: Chapman & Hall/CRC.

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

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, 39, 1–22.

Gravel, C. A. and R. W. Platt (2018): "Weighted estimation for confounded binary outcomes subject to misclassification," *Statistics in medicine*, 37, 425–436.

Greenland, S. (2000): "An introduction to instrumental variables for epidemiologists," *International journal of epidemiology*, 29, 722–729.

Hernán, M. and J. Robins (2020): *Causal Inference: What If*, Boca Raton: Chapman & Hall/CRC.

Holland, P. (1986): "Statistics in causal inference," *Journal of the American Statistical Association*, 81, 945–960.

Holland, P. (1988): "Causal inference, path analysis, and recursive structural equations models," *Sociological Methodology*, 18, 449–484.

Lipsitch, M., E. T. Tchetgen, and T. Cohen (2010): “Negative controls: a tool for detecting confounding and bias in observational studies,” *Epidemiology (Cambridge, Mass.)*, 21, 383.

Neyman, J., K. Iwaszkiewicz, and St. Kolodziejczyk (1935): “Statistical problems in agricultural experimentation,” *Supplement to the Journal of the Royal Statistical Society*, 2, 107–180.

Pearl, J. (2009): *Causality: Models, Reasoning and Inference*, New York: Cambridge University Press.

Pearl, J. (2010): “On the consistency rule in causal inference: Axiom, definition, assumption, or theorem?” *Epidemiology*, 21.

Petersen, M. L. and M. J. Van der Laan (2014): “Causal models and learning from data: integrating causal modeling and statistical estimation,” *Epidemiology (Cambridge, Mass.)*, 25, 418–426.

Robins, J. M., M. A. Hernan, and B. Brumback (2000): “Marginal structural models and causal inference in epidemiology,” *Epidemiology*, 11.

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

Rubin, D. (1974): “Estimating causal effects of treatments in randomized and nonrandomized studies,” *Journal of Educational Psychology*, 66, 688–701.

Rubin, D. B. (1976): “Inference and missing data,” *Biometrika*, 63, 581–592.

Rubin, D. B. (1987): *Multiple imputation for survey nonresponse*, New York: Wiley.

VanderWeele, T. J. (2009): “Concerning the consistency assumption in causal inference,” *Epidemiology*, 20, 880–883.

# 2

# A comparison between full text mining and searching in title, abstract and keywords for systematic reviews of epidemiological practice

Bas B. L. Penning de Vries
Maarten van Smeden
Frits R. Rosendaal
Rolf H. H. Groenwold

*Journal of Clinical Epidemiology* 2020; 121: 55–61

## Abstract

*Objective.* Article full texts are often inaccessible via the standard search engines of biomedical literature, such as PubMed and Embase, which are commonly used for systematic reviews. Excluding the full text bodies from a literature search may result in a small or selective subset of articles being included in the review because of the limited information that is available in only title, abstract and keywords. This article describes a comparison of search strategies based on a systematic literature review of all manuscripts published in 5 top-ranked epidemiology journals between 2000 and 2017. *Study Design and Setting.* Based on a text-mining approach, we studied whether 9 different methodological topics were mentioned across text fields (title, abstract, keywords, and text body). The following methodological topics were studied: propensity score methods, inverse probability weighting, marginal structural modelling, multiple imputation, Kaplan-Meier estimation, number needed to treat, measurement error, randomized controlled trial, and latent class analysis. *Results.* In total, 31,641 Hypertext Markup Language (HTML) files were downloaded from the journals' websites. For all methodological topics and journals, at most 50% of articles with a mention of a topic in the text body also mentioned the topic in the title, abstract or keywords. For each topic, a gradual decrease over calendar time was observed of reporting in the title, abstract or keywords. *Conclusion.* Literature searches based on title, abstract and keywords alone may not be sufficiently sensitive for studies of epidemiological research practice. This study also illustrates the potential value of full text literature searches, provided there is accessibility of full text bodies for literature searches.

## 2.1 Introduction

Rigorous reviews of the scientific literature are essential for determining the current state of knowledge on a specific topic, to identify research areas where evidence is lacking, and as a starting point for guidance development. While a majority of systematic reviews in epidemiology represents reviews of research findings on a specific substantive medical research topic, such as the occurrence of a particular disease or the effectiveness of a medical treatment, an important category of systematic reviews is concerned primarily with epidemiological research practice and reporting (Ali et al., 2015; Mendes and Batel-Marques, 2017; Brakenhoff et al., 2018; Copsey et al., 2018; Alfian et al., 2019).

A variety of strategies exist to identify and screen articles for eligibility for systematic reviews (Conn et al., 2003; O'Mara-Eves et al., 2015; Page et al., 2016; Lefebvre et al., 2018). Often, a staged search and screening approach is implemented in which the eligibility criteria for articles are made more stringent or more text fields are scrutinized with each step. In the earlier steps of the process, articles are typically excluded from the review on the basis of a small portion—e.g., title, abstract and keywords (TIABKW)—of all the available information. The goal of a search and screening approach is to identify all or a representative sample of the relevant literature on the topic of enquiry. However, excluding a selective set of articles from further study may ultimately result in a false impression of state of the literature being conveyed (O'Mara-Eves et al., 2015; Lefebvre et al., 2018; Egger and Smith, 1998).

Reviews of methods often begin searching for relevant literature in the same way as reviews on a substantive research topic. However, compared with substantive topics, the epidemiological and statistical methods used are likely less well documented in the small portion of information that is typically accessed in the first stage(s) of a systematic literature search, notably TIABKW. In this article, we investigate whether the traditional approach to systematic literature searching is appropriate for reviews of epidemiological practice.

## 2.2 Methods

We identified and downloaded all articles (in HTML format) published in the period 2000-2017 on the websites of five top-ranked epidemiological journals; Epidemiology (EPI), Journal of Clinical Epidemiology (JCE), European Journal of Epidemiology (EJE), International Journal of Epidemiology (IJE), and American Journal of Epidemiology (AJE).

All retrieved HTML pages were analyzed with R Statistical Software R Core Team (2018). First, we sought to extract for each article its publication date, title, abstract, keywords, and text body, in a largely automated fashion using R base regular expression algorithms (see e.g. Crawley, or Supplementary R Code). In-text references and reference lists were removed from the text bodies prior to analysis. The following methodological topics were selected for investigation: propensity score methods (PS), inverse probability weighting (IPW), marginal structural modelling (MSM), multiple imputation (MI), Kaplan-Meier estimation (KM), number needed to treat (NNT), measurement error (ME), randomized controlled trial (RCT), and latent class analysis (LC). This set of topics reflects a range of classical and modern methodological topics relevant to epidemiologic research. We subsequently determined for each of these topics whether there was any mention of the topic (see Supplementary Table S2.1 for details on the search terms) and in which text field (title, abstract, keywords, and text body).

The results of the previous step were used to quantify sensitivities of fixed combinations of text fields for identifying a mention of the method in any of the article's text fields (title, abstract, keywords or text body). For any fixed topic, we refer to the sensitivity of a particular combination of text fields (e.g., TIABKW) as the fraction of articles with a mention of the topic in any of these text fields among articles with a mention in the full text (i.e., in the title, abstract, keywords or body). We computed sensitivities stratified by journal and by publication date (year of publication). In a sensitivity analysis, the set of articles was limited to those articles containing at least 2500 words with the aim of focusing on original research articles. Additionally, we examined all articles with a mention of propensity score methods to determine the article type and whether or not the article described an empirical application of propensity score methods. Finally, we performed a post-hoc analysis, designed to ignore 'irrelevant' topic mentions (e.g., mention of a topic in the introduction or discussion of an article only). In this analysis, we considered only topics mentioned in the methods and results sections, provided these sections could be readily identified. Sensitivities pertaining to this post-hoc analysis are understood to refer to the fraction of articles with a mention of the topic in any of a given set of text fields among articles with a mention in the title, abstract, keywords, methods or results text fields.

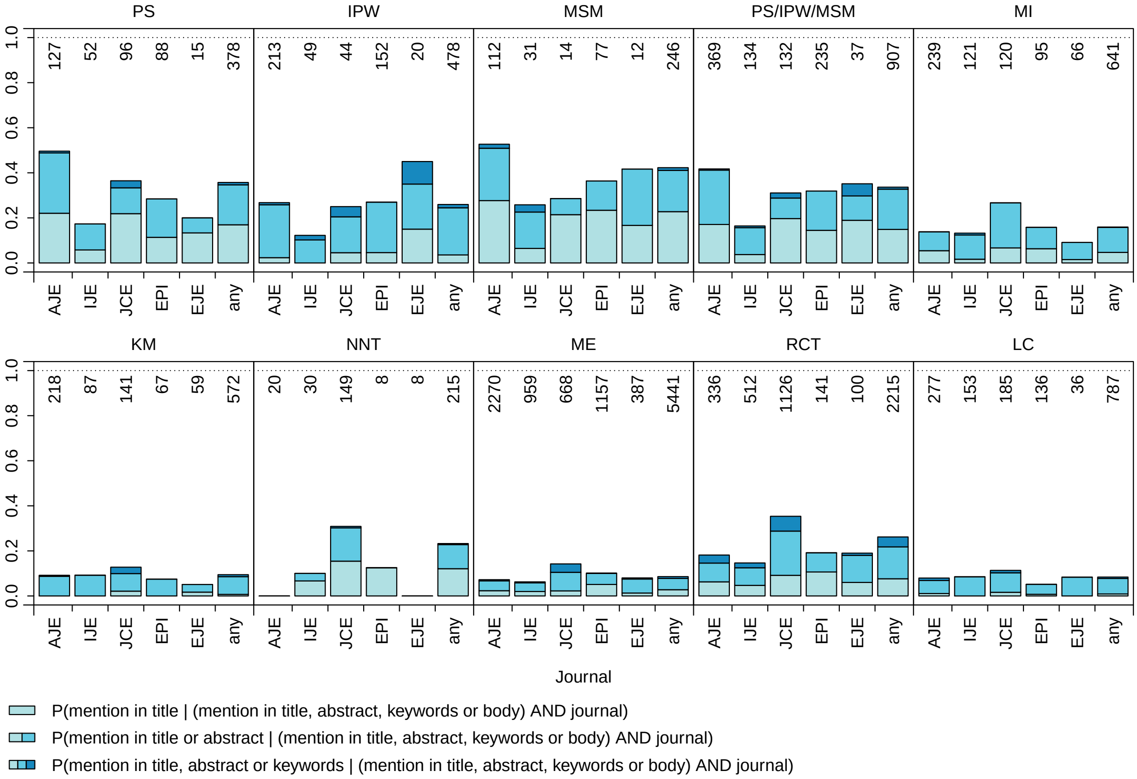


**Figure 2.1:** Sensitivities of topic mentioning in various text fields stratified by journal. Colors relate to text fields as follows: light blue areas give the proportion of articles with a topic mention in the title among all articles published in the indicated journal with a mention in the title, abstract, keywords, or body; light blue and blue areas together give the proportion of articles with a topic mention in the title or abstract; and light blue, blue, and dark blue areas together give the proportion of articles with a topic mention in the title, abstract, or keywords. PS, propensity score; IPW, inverse probability weighting; MSM, marginal structural modeling; MI, multiple imputation; KM, Kaplan-Meier; NNT, number needed to treat; ME, measurement error; RCT, randomized controlled trial; LC, latent class; AJE, American Journal of Epidemiology; IJE, International Journal of Epidemiology; JCE, Journal of Clinical Epidemiology; EPI, Epidemiology; EJE, European Journal of Epidemiology.

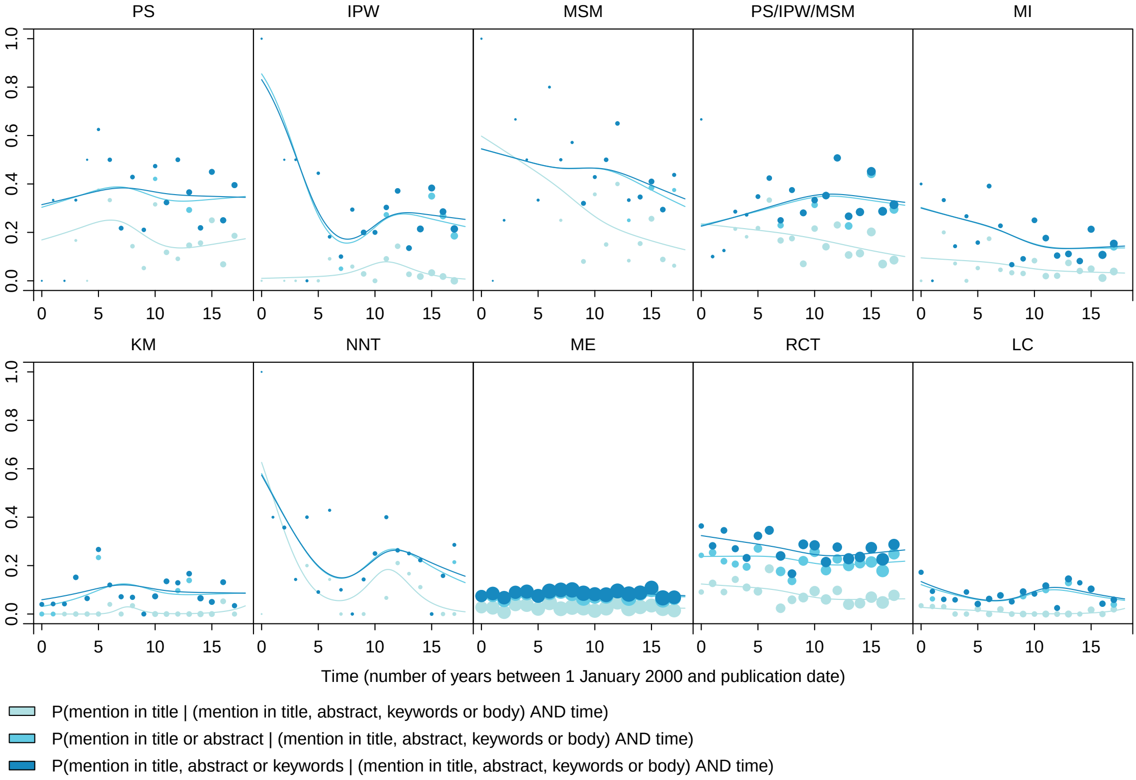


**Figure 2.2:** Sensitivities of topic mentioning in various text fields over time. Bullets give year-specific sensitivities with bullet size being proportional to number of publications in the given year with a mention of the topic in any text field (title, abstract, keywords, or body). Solid lines reflect logistic regression fits with cubic spline transformations of publication date with four knots placed equidistantly within [2000, 2017]. Colors relate to text fields as follows: for any given journal, light blue lines give the year-specific sensitivities of a topic mention in the title for a mention in the title, abstract, keywords, or body; blue lines indicate the year-specific sensitivities of a topic mention in the title or abstract;

and dark blue areas give the year-specific sensitivities of a topic mention in the title, abstract, or keywords. PS, propensity score; IPW, inverse probability weighting; MSM, marginal structural modeling; MI, multiple imputation; KM, Kaplan-Meier; NNT, number needed t o treat; ME, measurement error; RCT, randomized controlled trial; LC, latent class.

## 2.3 Results

We downloaded 31,641 HTML files from the journals' websites; 10,580 from EPI, 4,187 from JCE, 2,251 from EJE, 6,249 from IJE, and 8,374 from AJE. These files include (but are not limited to) what is published in HTML format of (indexed) articles, issue index pages and conference abstracts. Here, we present results based on those 31,641 files. In the Supplementary Material, results are presented on the subset of publications with at least 2500 words, for which results are comparable with what is presented here (Supplementary Figures S2.1 and S2.2).

Figures 2.1 and 2.2 present the sensitivities of TIABKW stratified by journal and by publication date, respectively. At most 50% of articles with a topic mention in any text field had a mention in the title, abstract or keywords. Figures 2.3 and 2.4 depict the results for our post-hoc analysis. For some topics (e.g., PS, MSM, and RCT), TIABKW mentions were considerably more sensitive for a topic mention in the full text excluding rather than including introduction and discussion. For other topics (e.g., MI, KM, and LC), TIABKW identified fewer than half the number articles with a topic mention anywhere in the full text, regardless of whether introduction and discussion were excluded. Some methodological topics had a constant, low, sensitivity throughout the study period (e.g., KM), whereas the sensitivity of TIABKW for the other topics gradually declined over time (e.g., MI, PS, IPW). There were no relevant differences in sensitivities of the reporting of topics across the different journals. Focusing on the articles that mention PS in the full text, 247 out of 378 articles mentioned PS in the text body but not in the title, abstract or keywords. Almost a third (72/247, 29%) of these described an empirical application of the method. This rate was more than doubled after we selected only those articles that, based on the nature of their main conclusion, were deemed predominantly applied research (60/87, 69%). Of the 131 articles that mentioned PS in the title, abstract or keywords, 82 (63%) described an empirical application. The positive predictive value of TIABKW for an empirical application was higher among predominantly empirical/applied original articles (58/60, 97%).

## 2.4 Discussion

Search engines that limit the searching of scientific articles to TIABKW, such as PubMed or Embase, are established starting points for systematic reviews of substantive epidemiological study questions (e.g., systematics reviews of the effects of a medical treatment). Our study illustrates that in systematic reviews of research practice and reporting, searches that rely only on these tools may lead

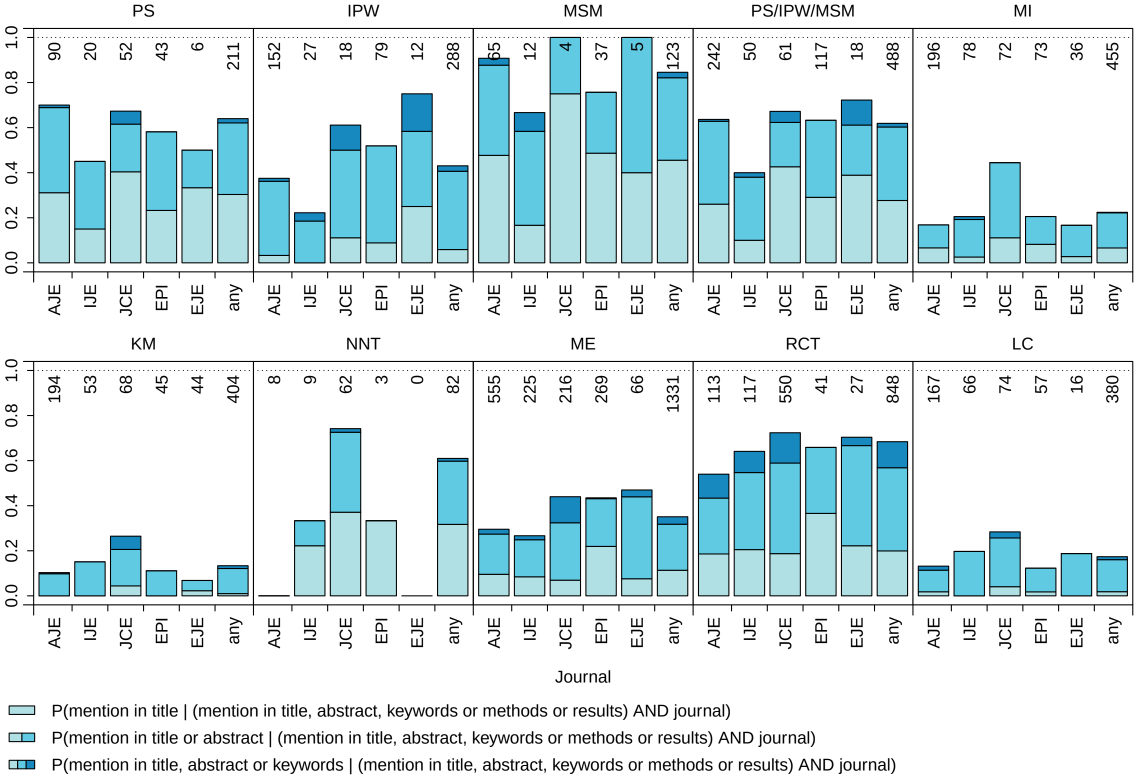


**Figure 2.3:** Sensitivities of topic mentioning in various text fields stratified by journal, according to post hoc analysis. Colors relate to text fields as follows: light blue areas give the proportion of articles with a topic mention in the title among all articles published in the indicated journal with a mention in the title, abstract, keywords, methods, or results text fields; light blue and blue areas together give the proportion of articles with a topic mention in the title or abstract; and light blue, blue, and dark blue areas together give the proportion of articles with a topic mention in the title, abstract, or keywords. PS, propensity score; IPW, inverse probability weighting; MSM, marginal structural modeling; MI, multiple imputation; KM, Kaplan-Meier; NNT, number needed to treat; ME, measurement error; RCT, randomized controlled trial; LC, latent class.

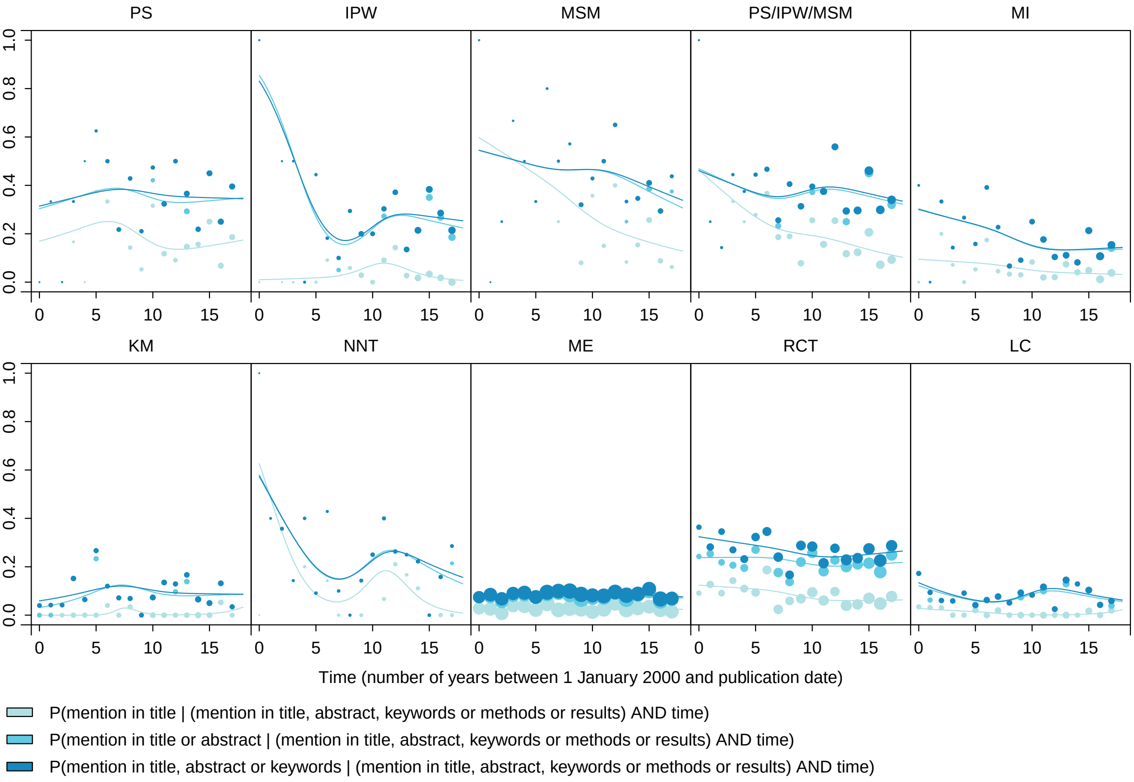


**Figure 2.4:** Sensitivities of topic mentioning in various text fields over time, according to post hoc analysis. Bullets give year-specific sensitivities for a mention in the title, abstract, keywords, methods, or results text fields, with bullet size being proportional to number of publications in the given year with a mention of the topic in title, abstract, keywords, or methods or results (provided the text field was identified and extracted). Solid lines reflect logistic regression fits with cubic spline transformations of publication date with four knots placed equidistantly within [2000, 2017]. Colors relate to text fields as follows: for any given journal, light blue lines give the year-specific sensitivities of a topic mention in the title for a mention in the title, abstract, keywords, or body; blue lines indicate the year-specific sensitivities of a topic mention in the title or abstract;

and dark blue areas give the year-specific sensitivities of a topic mention in the title, abstract, or keywords. PS, propensity score; IPW, inverse probability weighting; MSM, marginal structural modeling; MI, multiple imputation; KM, Kaplan-Meier; NNT, number needed to treat; ME, measurement error; RCT, randomized controlled trial; LC, latent class.

to a small and possibly selective subset of articles to be included in the review. We found a large discrepancy in terms of the number of articles identified (as potentially eligible) between searches that include text bodies and those that are restricted to title, abstract and keywords. Moreover, methodological topics tended to be documented in less detail in the title, abstract, or keywords as methods become more mainstream, contributing to a possibly selective subset of articles to be identified over time.

Reviewers are faced with the challenge of adequately handling increasingly large volumes of literature, and ignoring certain text fields may help mitigate this problem, but it may come at the cost of giving an inaccurate reflection of the state of knowledge/practice on the topic of interest. The decision to automate the selection of articles in systematic reviews using readily available search engines is usually made on practical grounds. Full text mining may however be a promising alternative. As noted by O'Mara-Eves et al., there are at least two (not necessarily distinct) ways of using data and text mining in selecting articles for further review: by reducing the list of items to be screened manually or by manually assigning articles in a (development) subset of articles to include/exclude categories in order to 'train' an algorithm to apply such categorizations automatically. Depending on the complexity of the task for which the algorithm is to be trained and the desired properties the trained algorithm should possess, the second (supervised-learning) approach may actually be more cumbersome than going through all articles manually. For the current analysis, we used text mining only to prune articles that would be deemed related to the topic of interest had we manually evaluated the paper. In some settings, e.g., where diverse or non-specific terminology is used, it may be difficult to find a rule that allows for relevant articles to be identified with high sensitivity and manageable specificity. In such cases, the adopted text mining approach may still leave an intractably large amount of articles to screen manually.

While our review clearly shows a possibly large difference between TIABKW searching versus full text searching, the discrepancies we found in this review in the number of pruned articles need not always translate into the two approaches giving a different impression of the state of research practice for any given methodological topic in epidemiology. This may depend on the review goals. Also, even if articles are missed by limiting the research to TIABKW, an important question remains whether the articles that would be omitted if we ignored the full text, should have actually been included. The large discrepancy that we found for the topic RCT, for example, is likely largely explained by many articles only briefly addressing the study design in the discussion or introduction, i.e., studies that may not be relevant to the reviewer (depending on the review goals) (see Figure 2.3). That is, incorporating all available text fields in the screening is likely to decrease the specificity for relevant articles, resulting in a possibly much larger number of articles to be further screened on relevance. It may therefore sometimes be appropriate to restrict oneself to certain text fields. Of note, for the topic of PS, many studies that would be omitted by restricting the search to TIABKW actually detailed an empirical application of the method. Therefore, for reviews of research practice regarding PS many relevant articles would be missed if the search/screening had been restricted to TIABKW only, especially the more recently published articles.

A limitation of this study is that it was limited to only five high ranking epidemiological journals and nine (partly related) methodological topics. Each of these journals has a strong methodological focus, publishing on applied as well as methodological topics. Consequently, we may expect that our results do not directly translate to other fields, particularly to applied biomedical journals with a less methodological focus.

There are several operational and legal challenges to consider for full automated text data literature searches. Clearly, if researchers do not have access to the full text of articles, initial screening based on title and abstract might be the only viable option. Furthermore, in case of hundreds of thousands of full text articles to be searched, downloading of the articles needs to be automated to, which is currently prohibited by some publishers. An alternative approach could be to restrict the search to open access articles only, but whether this is a suitable alternative depends on the objective of the review. Furthermore, there are practical barriers to perform full text searches, since this is not possible via commonly used search engines such as Pubmed.

Given the various challenges to automated searches, in current practice, there probably exists a trade-off between automated full-text literature searching in a small number of journals or TIABKW searching in large databases. Although not used in this study, both approaches could be supplemented with pearl growing strategies such as MeSH terms and snowballing in an effort to increase the sensitivity (Greenhalgh and Peacock, 2005; Ramer, 2005).

To conclude, searches that are based on TIABKW only may not be appropriate for systematic reviews of research practice and reporting. Provided access to full text bodies for literature searches, full text mining is ideally incorporated also in the first stages of a systematic literature review of epidemiological practice.

## References

S. D. Alfian, I. S. Pradipta, E. Hak, and P. Denig (2019): "A systematic review finds inconsistency in the measures used to estimate adherence and persistence to multiple cardiometabolic medications," *Journal of Clinical Epidemiology*, 108, 44–53.

Ali, M. S., R. H. Groenwold, S. V. Belitser, W. R. Pestman, A. W. Hoes, K. C. Roes, A. de Boer, and O. H. Klungel (2015): "Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review," *Journal of Clinical Epidemiology*, 68, 122–131.

Brakenhoff, T. B., M. Mitroiu, R. H. Keogh, K. G. Moons, R. H. Groenwold, and M. van Smeden (2018): "Measurement error is often neglected in medical literature: a systematic review," *Journal of Clinical Epidemiology*, 98, 89–97.

V. S. Conn, S.-A. Isaramalai, S. Rath, P. Jantarakupt, R. Wadhawan, and Y. Dash (2003): "Beyond MEDLINE for literature searches," *Journal of Nursing Scholarship*, 35, 177–182.

Copsey, B., J. Thompson, K. Vadher, U. Ali, S. Dutton, R. Fitzpatrick, S.E. Lamb, and J.A. Cook (2018): "Sample size calculations are poorly conducted and reported in many randomised trials of hip and knee osteoarthritis: results of a systematic review," *Journal of Clinical Epidemiology*, 105, 52–61.

Crawley, M. J. (2013): "2.12: Text characters strings and pattern matching." In: M. J. Crawley [editor], *The R Book*, Chichester: John Wiley & Sons.

Egger, M., and G. D. Smith (1998): "Meta-analysis bias in location and selection of studies," *BMJ*, 326, 61–66.

Greenhalgh, T., and R. Peacock (2005): "Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources," *BMJ*, 331, 1064–1065.

Lefebvre, C., E. Manheimer, and J. Glanville (2018): "Chapter 6: Searching for studies." In: J. P. T. Higgins, S. Green [editors], *Cochrane Handbook for Systematic Reviews of Interventions: Version 5.1.0*, Oxford: The Cochrane Collaboration.

Mendes, D., Alves, C., and F. Batel-Marques (2017): "Number needed to treat (NNT) in clinical literature: an appraisal," *BMC Medicine*, 15, 112.

O'Mara-Eves, A., J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou (2015): "Using text mining for study identification in systematic reviews: a systematic review of current approaches," *Systematic Reviews*, 4, 5.

M. J. Page, L. Shamseer, D. G. Altman, J. Tetzlaff, M. Sampson, A. C. Tricco, F. Catalá-López, L. Li, E. K. Reid, R. Sarkis-Onofre, et al. (2016): "Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study," *PLoS medicine*, 13, e1002028.

Ramer, S. L. (2005): "Site-ation pearl growing: methods and librarianship history and theory," *Journal of the Medical Library Association*, 93, 397.

R Core Team (2018): *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `https://www.R-project.org/`.

## Supplementary Material

| Methodological topic | Search terms |
|---|---|
| Methodological topic | Search terms |
| Propensity score methods | propensity score; propensity scoring |
| Inverse probability weighting | “inverse-probability-weight; inverse probability-weight; inverse-probability weight; inverse probability weight; inverse weight; inverse-weight; inverse-probability; inverse probability” |
| Marginal structural modelling | marginal structural |
| Multiple imputation | multiple imputation; multiply imputed |
| Kaplan-Meier estimation | kaplan-meier; kaplan meier |
| Number needed to treat | number needed to; number-needed-to |
| Measurement Error | misclassification; measurement error |
| Randomised controlled trial | randomized controlled clinical trial; randomised controlled clinical trial; randomized controlled trial; randomised controlled trial |
| Latent class analysis | latent variable; finite mixture |

**Table S2.1:** Overview of search terms (strings) for each of the methodological topics. Distinct strings are separated by semicolons. Articles with a mention of at least one of the specified search strings were regarded as eligible for review (i.e., as articles potentially referring to the topic of interest). A mention of a term/string was established using case insensitive approximate string matching with unit edit costs; an approximate match was said to exist if and only if the Levenshtein distance was no greater than 10% of the number of search term characters (i.e., based on the default settings of the R/3.5.0 function base::agrep).

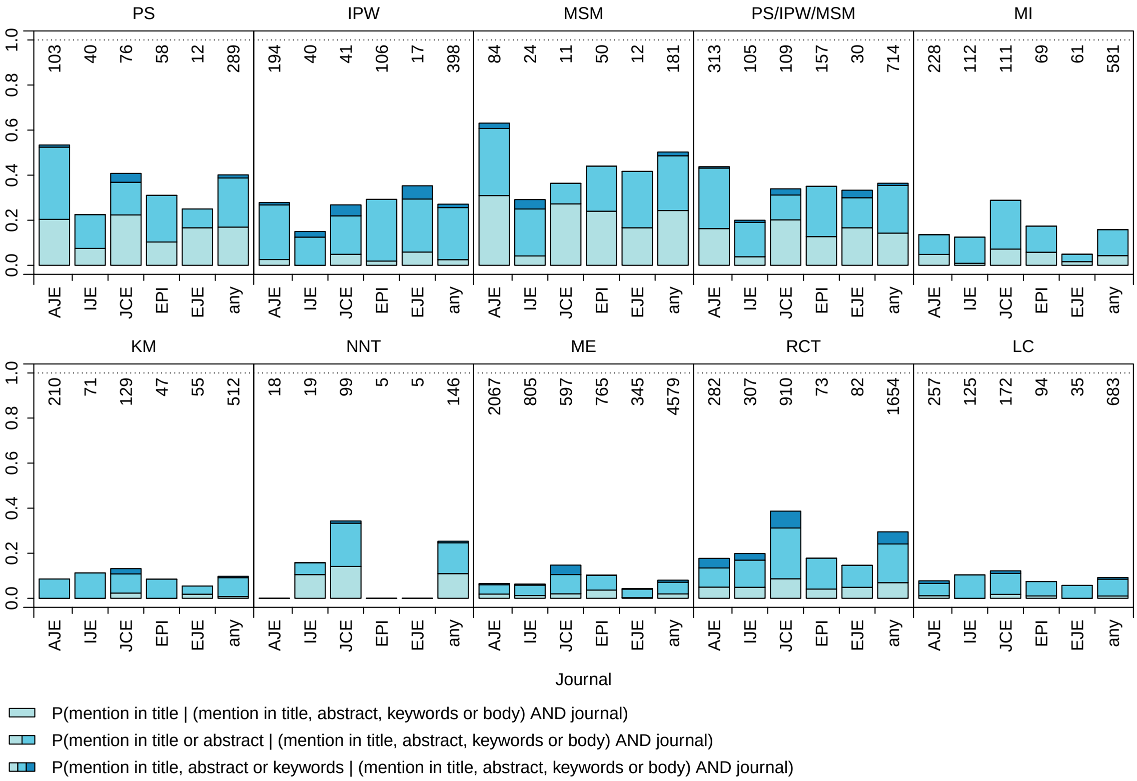


**Figure S2.1:** Sensitivities of topic mentioning in various text fields stratified by journal among articles of at least 2500 words.

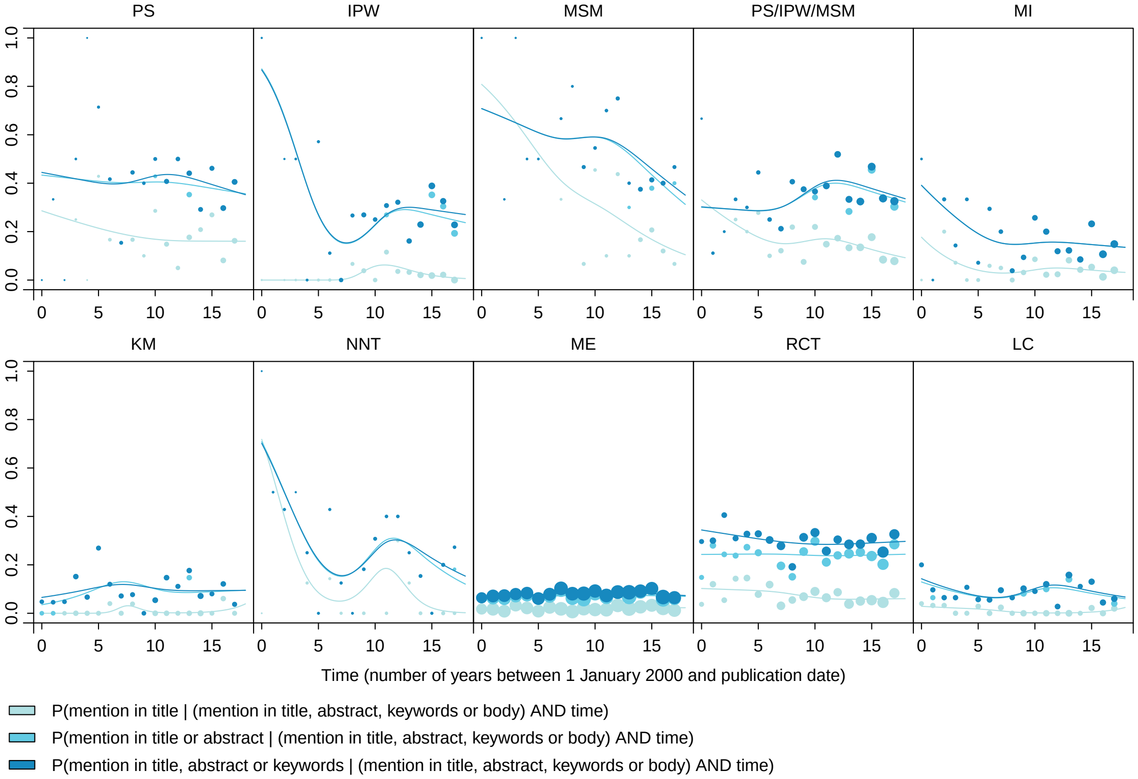


**Figure S2.2:** Sensitivities of topic mentioning in various text fields over time among articles of at least 2500 words. Bullets give year-specific sensitivities with bullet size being proportional to number of publications of at least 2500 words in the given year with a mention of the topic in any text field (title, abstract, keywords or body). Solid lines reflect logistic regression fits with cubic spline transformations of publication date with four knots placed equidistantly within [2000, 2017].

Chapter 2

# Supplementary R Code

```
# The R code below was compiled to illustrate how the page source of
#   the following article may be downloaded, how relevant parts may
#   be extracted or modified, and how term mentions may be
#   identified.

# Mi, X., B. G. Hammill, L. H. Curtis, M. A. Greiner, S. and
#   Setoguchi, 2013. Impact of immortal person-time and time scale
#   in comparative effectiveness research for medical devices: a case
#   for implantable cardioverter-defibrillators, Journal of clinical
#   epidemiology, 66(8), pp.S138-S144.

# ==================================================================
# Downloading file
# ==================================================================

# NB: The method used below retrieves the original/unrendered page
#  source. What is returned may not contain all elements that are
#  displayed by the web browser. Publisher APIs or headless browsers
#  may be helpful when this occurs. For this example, the original
#  page source is sufficient.

con <- url("https://www.jclinepi.com/article/S0895-4356(13)00163-7/
  fulltext",method="libcurl")
x <- suppressWarnings(paste0(readLines(con),collapse=" \n "))

# ==================================================================
# Extracting relevant parts
# ==================================================================

# Page sources from the same journal typically have the same
#   structure. Inspection of some source files should help with
#   locating and, in turn, extracting the relevant parts. Below we
#   make use of regular expressions.
?regexpr # But see also
# Crawley, M. J. (2013). 2.12: Text characters strings and pattern
#   matching. In M. J. Crawley [editor]. The R Book, Chichester: John
#   Wiley & Sons.

# Title
m <- regexpr('<h1 class=\"articleTitle\">(.*?)</h1>',x)
title <- regmatches(x,m)

# Abstract
# To isolate the abstract we would like to extract everything
#   between '<h2 class="<section class="abstract'
#   and the matching '</section>'. However, '</section>'
#   occurs multiple times in this target string, so greedily
#   searching for '</section>' after
#   '<h2 class="<section class="abstract' is not effective here.
#   The following functions handle this problem.

fMatchOpenClose <- function(open='<div',close='</div>',text){
  opn <- gregexpr(open,text)[[1L]]
  cls <- gregexpr(close,text)[[1L]]
  n <- length(opn)
  if(n!=length(cls))
    stop(paste('discrepancy between number of opening and',
      'that of closing strings.')
  wh <- integer(n)
  available <- !wh
  for(i in seq_len(n)){
    sgn <- c(rep(1L,n-i+1),rep(-1,sum(available)))
    loc <- c(opn[i:n],cls[available])
    ord <- order(loc)
    sgn <- sgn[ord]
    loc <- loc[ord]
    clr <- loc[!cumsum(sgn)]
    if(!length(clr)) stop('invalid entry.')
    wh[i] <- clr[1L]
    available[which(cls==clr[1L]&available)[1L]] <- FALSE
  }
  m <- as.integer(opn)
  attr(m,'match.length') <- as.integer(wh+nchar(close)-opn)
  m <- list(m)
  return(m)
}
fMatchOpenCloseStartStop <- function(
  start='<div id="fulltext-body">',
  open='<div',close='</div>',stop='</div>',text){
  m <- fMatchOpenClose(open,close,text)[[1L]]
  str <- as.integer(regexpr(start,text))
  stp <- as.integer(gregexpr(stop,text)[[1L]]+nchar(stop))-1L
  m0 <- as.integer(m)
  m1 <- as.integer(m+attr(m,'match.length')-1L)
  wh <- m0>=str
  m0 <- m0[wh]
  m1 <- m1[wh]
  w <- rep(seq_len(length(m0)),2L)
  s <- c(rep(1L,length(m0)),rep(-1L,length(m1)))
  m <- c(m0,m1)
  o <- order(m)
  w <- w[o][which(!cumsum(s[o]))[1L]]
  stp <- stp[which(stp>=m1[w])[1L]]
  attr(str,'match.length') <- stp-str+1L
  return(str)
}
m <- fMatchOpenCloseStartStop(start='<section class=\"abstract',
  open='<section',close='</section>',stop='</section>',text=x)
abstract <- regmatches(x,m)

# Keywords
m <- regexpr('<div class=\"keywords\">(.*?)</div>',x)
keywords <- regmatches(x,m)

# Body
m <- fMatchOpenCloseStartStop(
  start='<div class=\"content\"><section id="sec1"',open='<div',
  close=’</div>’,stop=’</div>’,text=x)
body <- regmatches(x,m)
# NB: the reference list is not included in this text field.

# ==================================================================
# Modifying/cleaning extracted parts
# ==================================================================

title <- gsub("<(.*?)>","",title)
abstract <- trimws(gsub("^(.*?)Abstract","",gsub("<(.*?)>"," ",abstract)))
keywords <- gsub("<(.*?)>","",keywords)
keywords <- trimws(gsub("Keywords: \n","",keywords))
unlist(strsplit(keywords,", ")) # note that "Defibrillators" and
# "implantable" now appear as distinct keywords because of the crude
#   approach.
body <- gsub(
  paste0(’\\[<span class="bibRef\"(.*?)’,
     ’See all References</a></span></span>\\]’),"",body)
  # This also removes in-text references.
body <- gsub("<(.*?)>"," ",body)

# ==================================================================
# Partial string matching
# ==================================================================

agrepl("Inverse probability weighting",title,ignore.case=TRUE) #F
agrepl("Inverse probability weighting",abstract,ignore.case=TRUE) #F
agrepl("Inverse probability weighting",keywords,ignore.case=TRUE) #F
agrepl("Inverse probability weighting",body,ignore.case=TRUE) #T

# The following function may be used to extract parts where a partial
#   string match is found. A measure of the location of the match is
#   also given.
getExcerpts <- function(term,before=’. ’,after=’. ’,count=1L,
  fixed=TRUE,text,min_dist=25L,trim_start=before,trim_end=NULL){
  if(count<1L||!is.integer(count)) stop("count must be positive integer.")
  if(min_dist<1L||!is.integer(min_dist))
    stop("else_nchar must be positive integer.")
  x <- gregexpr(term,text,ignore.case=TRUE)[[1L]]
  y <- as.integer(x)+attr(x,"match.length")-1L
  x <- as.integer(x)
  a <- gregexpr(before,text,fixed=fixed)[[1L]]
  b <- as.integer(a)+attr(a,"match.length")-1L
  c <- gregexpr(after,text,fixed=fixed)[[1L]]
  d <- as.integer(c)+attr(c,"match.length")-1L
  n <- nchar(text)
  fn <- function(i){
    v <- rev(a[b<x[i]])
    m <- min(c(length(v),count))
    p <- if(!is.na(x[i])&&m>0&&x[i]-v[m]>=min_dist) v[m] else
      max(c(1,x[i]-min_dist))
    w <- d[c>y[i]]
    m <- min(c(length(w),count))
    q <- if(!is.na(x[i])&&m>0&&w[m]-x[i]>=min_dist) w[m] else
      min(c(n,x[i]+min_dist))
    z <- p
    attr(z,"match.length") <- q-p+1L
    out <- regmatches(text,z)
    if(!is.null(trim_start))
      out <- gsub(paste0('^',trim_start),'',out)
    if(!is.null(trim_end))
      out <- gsub(paste0(trim_end,'$'),'',out)
    return(list(excerpt=trimws(out),location=x[i]/n))
  }
  l <- length(x)
  out <- if(l) lapply(seq_len(l),fn) else NA
  out <- list(excerpt=unlist(lapply(out,function(x)x$excerpt)),
    location=unlist(lapply(out,function(x)x$location)))
  return(out)
}
getExcerpts("inverse probability",text=body)
```

# 3

# Comments on propensity score matching following multiple imputation

Bas B. L. Penning de Vries
Rolf H. H. Groenwold

*Statistical Methods in Medical Research* 2016; 25(6): 3066–3068

In a recently published simulation study, Mitra and Reiter compared two approaches to implementing propensity score (PS) methods following multiple imputation (Mitra and Reiter, 2016). Particular emphasis was on propensity score matching following multiple imputation. In simulation studies, they evaluated two possible approaches, i.e., the so-called Within and Across approach. In both approaches, PSs are estimated in each of $m$ imputed datasets. In the Within approach, PS matching is performed within each imputed dataset. The resulting $m$ effect estimates are then pooled by averaging. In the Across approach, for each subject the $m$ estimated PSs are averaged first, after which PS matching is performed once, based on each subject's average PS. Apparent from the results was the trend that although both approaches were biased, the Within method was generally more biased than the Across approach, particularly when there was missing confounder data.

We argue that these findings are due to the imputation model and the matching algorithm rather than a genuine difference between the methods. While Mitra and Reiter chose to leave the outcome out of the imputation model, it has been shown that often the outcome should actually be included in the imputation model (van Buuren, 2012). To illustrate this, we repeated a selection of Mitra and Reiter's simulations, which represent a setting of a binary treatment, a continuous outcome, and two normally distributed covariates. Here, we focus on the scenarios in which both covariates acted as confounders and both treated and untreated subjects were assigned missing covariate values. Results are presented in Table 3.1.

In line with Mitra and Reiter, when we applied PS matching while leaving the outcome variable out of the imputation model, the Across approach outperformed the Within approach in terms of bias (Table 1, scenario 1a). With the outcome included in the imputation model (scenario 1b), the Within estimates still deviate more from the true treatment effect than the Across estimates, but closely approximate the mean estimate based on PS matching before the introduction of missing values (0.053, 95%CI 0.043; 0.064). The bias observed in the absence of missing data is largely due to non-positivity in the tails of the PS distributions of treated and untreated subjects. As a result, treated subjects in the upper tail of the PS distribution are matched to untreated subjects who tend to have lower PS values, thus leading to suboptimal balance in PS between treated and untreated subjects. This balance can, however, be improved, e.g., by using narrow callipers for matching or increasing the sample size $n$ and thus increasing the number of potential matches. With $n$ increased by a factor of 10 (Table 1, 1c) it becomes apparent that the Within approach is superior to the Across provided the outcome variable is included in the imputation model.

| | Across | | | Within | | |
|---|---|---|---|---|---|---|
| | Pt. Est. | Variance | MSE | Pt. Est. | Variance | MSE |
| Scenario 1: matching | | | | | | |
| a | 0.282 | 0.081 | 0.161 | 0.557 | 0.051 | 0.361 |
| b | -0.012 | 0.032 | 0.032 | 0.060 | 0.024 | 0.027 |
| c | -0.048 | 0.003 | 0.005 | 0.025 | 0.002 | 0.003 |
| Scenario 2: regression | | | | | | |
| a | 0.166 | 0.039 | 0.066 | 0.438 | 0.032 | 0.224 |
| b | -0.077 | 0.020 | 0.026 | 0.002 | 0.018 | 0.018 |
| Scenario 3: IPTW | | | | | | |
| a | 0.092 | 0.002 | 0.010 | 0.043 | 0.001 | 0.003 |
| b | -0.701 | 0.805 | 1.296 | 0.227 | 0.616 | 0.668 |
| c | 0.011 | 0.001 | 0.002 | -0.002 | 0.001 | 0.001 |
| d | -0.236 | 0.664 | 0.720 | -0.022 | 0.658 | 0.658 |
| e | 0.901 | 0.009 | 82.793 | 0.888 | 0.009 | 83.030 |
| f | 9.638 | 2.008 | 2.139 | 9.967 | 2.065 | 2.066 |

**Table 3.1:** Properties of the Across and Within estimators. Abbreviations: Pt. Est. = mean effect estimate across 1000 simulations; Variance = empirical variance; MSE = mean squared error. Sample size $n = 1100$, except for scenario 1c where $n = 11,000$. In scenario 1, effect estimates were based on PS matching following multiple imputation with outcome left out (a) or included (b, c) in the imputation model. In scenario 2 (a, b), treatment effects were estimated using linear regression, regressing the outcome on treatment, the PS, and both covariates. In 2a, the outcome was left out of the imputation model, whereas in 2b it was included. In scenario 3, effect estimates were based on IPTW, following Mitra and Reiter (2016) (a, c, e), or using the traditional weights (see text) (b, d, f). In scenarios 3a and 3b the outcome variable was not included in the imputation model, whereas in scenarios 3c, 3d, 3e and 3f the outcome was included in the imputation model. In all scenarios, the true effect of treatment on the outcome was zero, except for scenarios 3e and 3f, in which it was 10.

Mitra and Reiter also assessed multiple imputation followed by regression adjustment (i.e., including the confounders as covariates in a linear regression model). In this situation, we observed the same trend across the different imputation models (Table 1, scenarios 2a and 2b). Again, upon inclusion of the outcome variable in the imputation model, the Within approach yields unbiased estimates, while the Across approach does not.

A third method of controlling for confounding that was studied by Mitra and Reiter was inverse probability weighting. In scenario 3a, we estimated the true effect using inverse probability of treatment weighting (IPTW) where, following Mitra and Reiter (2016), the weight for any subject equalled 1 if a subject was treated, and PS/(1-PS) if untreated. The treatment effect is then estimated by the difference between the sum of the weighted outcomes in the treatment group and the sum of the weighted outcomes in the control group, divided by the original sample size $n$. In scenario 3b, we used the traditional weights discussed by, e.g., Lunceford and Davidian (2004) and Robins et al. (2000); i.e., 1/PS if a subject was treated, and 1/(1-PS) otherwise. Note that these are equivalent to those of scenario 3a multiplied by PS, meaning that the average of weighted outcomes based on the weights used by Mitra and Reiter is necessarily closer to zero (since PS $<$ 1), than if the traditional weights were used. Again, we observed that with the outcome variable included in the imputation model, the Within method is superior to the Across (Table 1, scenarios 3c and 3d). Further, in scenarios of a non-null treatment effect (Table 1, 3e and 3f, true effect $=$ 10) simulations suggest that the traditional weights are to be preferred—i.e., unless the interest lies in estimating the average effect on the treated (Morgan and Todd, 2008), in which case the denominator of the effect estimator should match the effective size of the groups in the pseudopopulation.

In medical research, confounding and missing data are common problems that often occur simultaneously. When multiple imputation is to be followed by PS matching, researchers could apply the Across and the Within approaches that were proposed by Mitra and Reiter. Provided the correct imputation model is applied and there are no other sources of bias (e.g., model misspecification), the Within approach appears to be superior to the Across approach in terms of bias reduction.

## References

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

Mitra, R. and J. P. Reiter (2016): "A comparison of two methods of estimating propensity scores after multiple imputation," *Statistical methods in medical research*, 25, 188–204.

Morgan, S. L. and J. J. Todd (2008): "A diagnostic routine for the detection of consequential heterogeneity of causal effects," *Sociological Methodology*, 38, 231–281.

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

van Buuren, S. (2012): *Flexible imputation of missing data*, CRC Press.

# 4

# A comparison of two approaches to implementing propensity score methods following multiple imputation

Bas B. L. Penning de Vries
Rolf H. H. Groenwold

*Epidemiology, Biostatistics and Public Health* 2017; 14(4): e12630

## Abstract

*Background.* In observational research on causal effects, missing data and confounding are very common problems. Multiple imputation and propensity score methods have gained increasing interest as methods to deal with these, but despite their popularity methodologists have mainly focused on how they perform in isolation. *Methods.* We studied two approaches to implementing propensity score methods following multiple imputation, both of which have been used in applied research, and compared their performance by way of Monte Carlo simulation for a continuous outcome and partially unobserved covariate, treatment or outcome data. In the first, propensity score analysis is performed within each of $m$ imputed datasets, and the resulting $m$ effect estimates are averaged. In the alternative approach, for each subject the $m$ estimated propensity scores are averaged first, after which the propensity score method is implemented based on each subject's average propensity score. *Results.* The Within approach was found to be superior in terms of bias as well as variance in settings with missing covariate data. In settings with incomplete treatment or outcome values only, the approaches yielded similar results. Reasons for discrepancies between the approaches are provided. *Conclusion.* We advise researchers not to use the second approach as the default method, because even when data are missing completely at random, this may yield biased effect estimates.

Chapter 4

## 4.1 Introduction

Establishing causal associations between risk factors, treatments, or other exposures, and outcomes is a key aim in many epidemiologic studies. However, in observational studies, the attempt is often hampered by missing data and confounding.

Simply ‘ignoring’ missing data, which typically means that a complete case analysis is performed, often is inappropriate, because the conditions under which it is unbiased are very restrictive (Rubin, 2004; Schafer and Graham, 2002; van Buuren, 2012; Daniel et al., 2012). Even when these conditions are met, for example because the complete cases are truly a random subset of the study sample, discarding incomplete records may render the estimator unnecessarily inefficient (Rubin, 2004; Schafer and Graham, 2002; van Buuren, 2012). An alternative to complete case analysis is multiple imputation (MI), in which missing data are filled in with random draws from their predictive distributions based on the observed data, thereby producing multiple plausible datasets. Inferences are typically made using a simple set of rules known as Rubin’s Rules (Rubin, 2004). MI is a popular method for dealing with missing data, because it is flexible, relatively easy to implement with readily available statistical packages, and often provides valid estimates of effect and standard errors in situations where simpler techniques, including complete case analysis, fail (Rubin, 2004; Schafer and Graham, 2002; van Buuren, 2012).

To address the problem of confounding, researchers have traditionally used multivariable regression for data analysis. More recently, the use of propensity score methods has gained increasing interest (Stürmer et al., 2006). A subject’s propensity score is the conditional probability of being assigned to treatment given their measured covariates (Rosenbaum and Rubin, 1983; Austin, 2011). Among those subjects with the same propensity score, the distribution of measured covariates is expected to be the same between treated and untreated individuals (Rosenbaum and Rubin, 1983). Thus, by conditioning on the propensity score treatment status becomes independent of covariates. Several propensity score methods have been described: stratification, matching on the propensity score, inverse probability of treatment weighting (IPW), and covariate adjustment in multivariable regression (Rosenbaum and Rubin, 1983; Austin, 2011). However, despite increasing popularity, it is largely unclear how these perform in the presence of missing data.

Few have investigated approaches that combine missing data techniques with methods for confounding. Mitra and Reiter (2016) studied two approaches that combine multiple imputation with propensity score matching. In both, missing covariate data are imputed $m$ times through multiple imputation. For each of the completed datasets, a propensity score is then estimated for each subject. In the so-called Within approach, propensity score analysis is performed within each of $m$ imputed datasets, and the resulting $m$ effect estimates are averaged. In the Across approach, for each subject the $m$ estimated propensity scores are averaged first, after which the propensity score method is implemented based on each subject's average propensity score. While both approaches were shown to be superior to complete case analysis in terms of bias, it was found that the Across method was less biased than the Within method, especially in the presence of missing covariate data (Mitra and Reiter, 2016). However, as with any simulation study, these results may not extend beyond the settings that were considered. For example, while it was assumed that the treatment and outcome variables were fully observed, none have compared approaches in settings with incomplete data for one or both of these variables. Furthermore, although it has been argued that often the outcome should be included in the imputation model (Moons et al., 2006; Sterne et al., 2009; van Buuren, 2012), it was excluded from the imputation model in the previous study. Moreover, with the outcome included, subsequent simulations have found the Within approach to be preferred (Penning de Vries and Groenwold, 2016; Leyrat et al., 2017). Nevertheless, in applied research, the Across approach appears to have gained interest since its introduction (Neuderth et al., 2016; Olszewski et al., 2015; Gregory et al., 2016; Chiu et al., 2016; Brown et al., 2016; Sulkowski et al., 2014, 2015; Kutney-Lee et al., 2014; Ekström et al., 2016).

Our aim was therefore to provide further insight into how propensity scores analysis should be applied in combination with multiple imputation. Specifically, we compared the Within and Across approaches in settings with missing covariate data, missing treatment indicators, and missing outcomes. The remainder of this article is structured as follows. The notation and set-up for the simulations are detailed in Sections 4.2 and 4.3. Results are presented in Section 4.4 and discussed in Section 5.5. Finally, we conclude with a summary in Section 4.6.

## 4.2 Notation

Suppose the random vector $\boldsymbol{Z} = (X_1, X_2, ..., X_g, T, Y)$ is observed on $n$ subjects. The first $g$ variables of $\boldsymbol{Z}$ represent covariates, whereas $T$ and $Y$ refer to a binary treatment indicator variable and a continuous outcome, respectively. Realisations are printed in lower case letters. We denote an $n \times (g+2)$ matrix by $\mathbf{Z}$, whose $i$th row $\boldsymbol{Z}_i = (X_{i1}, X_{i2}, ..., X_{ig}, T_i, Y_i)$ represents the $i$th $(i = 1, 2, ..., n)$ subject's record. For each $i, j$ element in $\mathbf{Z}$, define a missing indicator variable $M_{ij}$ that takes the value of 1 if it is observed and 0 otherwise. Further, we write $\mathbf{z} = (\mathbf{z}^{\text{obs}}, \mathbf{z}^{\text{mis}})$ to indicate that $\mathbf{z}$ can be partitioned into an observed part $\mathbf{z}^{\text{obs}}$ and a missing part $\mathbf{z}^{\text{mis}}$. In multiple imputation, values of $\mathbf{z}^{\text{mis}}$ are imputed $m$ times by drawing from posterior predictive distributions, resulting in $m$ completed datasets $\mathbf{z}^{(k)}$, $k = 1, 2, ..., m$ that may be subjected to propensity score analysis. A detailed description of the Across and Within approaches are given in the Supplementary Material.

## 4.3 Simulation methods

We used a series of Monte Carlo simulations to examine the performance of the Across, the Within, and complete case approaches under various missing data mechanisms. The simulations were carried out in several stages. In the first stage, complete data are generated following one of the data generating mechanisms detailed below. These were chosen for comparability with Mitra and Reiter (2016). Second, missing data are introduced into one of the variables. Third, a number of approaches are applied to estimate the treatment-outcome effect. For each scenario (combination of complete data generating mechanism and missing data mechanism), this process was repeated 1000 times. A full factorial design was used. All simulations were conducted with R Statistical SoftwareR Core Team (2016) version 3.1.1. For multiple imputation we used the `mice` package (van Buuren and Groothuis-Oudshoorn, 2011). Continuous and binary variables were imputed using the `norm` and `logreg` options, respectively. The number of imputations was set to $m = 5$ for efficiency. For any incomplete variable, all other variables, including the outcome, were included in the imputation model. Failing to include the outcome in the imputation model may lead to imputed datasets that do not reflect the association between covariate and outcome that would have been observed had there been no missing values. The consequence of this is that if one adjusts for the imputed covariate values to estimate the treatment effect, the variation in outcomes between the treatment groups that is due to the partially unobserved covariate would in part be attributed to the differences in treatment status.

### *4.3.1 Data generating mechanisms*

We considered $g = 2$ covariates, a binary treatment indicator variable ($T_i$) and a continuous outcome for $n = 1100$ subjects. Data were simulated by sequentially drawing $(X_{i1}, X_{i2})$, $T_i$, and $Y_i$ for $i = 1, 2, ..., n$ from the respective distributions.

Let $(X_{i1}, X_{i2}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (10, 10)$ and $\boldsymbol{\Sigma}$ has variances equal to 5 and covariance 2.5 (correlation 0.5).

The value of $T_i$ was assigned by drawing from a Bernoulli distribution with parameter (i.e. the probability of treatment assignment) defined as a function of the $i$th subject's covariate data. In particular, we let

$$\Pr(T_i = 1|X_{i1}, X_{i2}) = \text{expit}\{-7.8 + 0.255X_{i1} + 0.255X_{i2}\}$$

where expit$\{\eta\}$ is the inverse logit function $\exp(\eta)/(1+\exp(\eta))$. As such, the log odds of treatment increases with 0.255 for every unit increase in either $X_1$ or $X_2$. This mechanism assigns approximately 100 subjects to treatment ($T = 1$) and 1000 subjects to the control group ($T = 0$).

We defined the outcome $Y_i$ such that, for all $i$,

$$Y_i = 2T_i + X_{i1} + 0.5X_{i2} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

where $\varepsilon_i$ is independent of $(T_i, X_{i1}, X_{i2})$. The interest lies in estimating treatment effect $\beta_{TY} = 2$, that is, the conditional treatment effect, which—because of homogeneity and the collapsibility of the causal difference in means—equals the marginal treatment effect. Clearly, both covariates serve as a confounder for the association between $T$ and $Y$. We varied $\sigma^2 = 1, 9$ to show that larger residual variances ($\sigma^2 = 9$) correspond with larger discrepancies between Across and Within estimates in the case of missing covariate data.

### *4.3.2 Missing data mechanisms*

To aid understanding, we initially restricted ourselves to simple missing data mechanisms, namely univariate missing completely at random (MCAR) mechanisms, and finally considered univariate missing at random (MAR) settings. The mechanisms for generating missing data were as follows:

*(i) MCAR covariate values.* Any subject's $X_2$ value was allowed to be missing with probability $\Pr(M_{i2} = 1|\mathbf{Z}) = p$, $p = 0.2, 0.4, 0.6, 0.8$. For columns $j = 1, 3, 4$ in $\mathbf{Z}$, we let $\Pr(M_{ij} = 1|\mathbf{Z}) = 0$.

*(ii) MCAR treatment indicator values.* We allowed for missing treatment status with $\Pr(M_{i3} = 1|\mathbf{Z}) = p$, $p = 0.2, 0.4, 0.6, 0.8$, and let the missingness probability of the other variables equal 0.

*(iii) MCAR outcome values.* We considered the same missing data mechanisms as in *ii* except that we simulated missing outcomes as opposed to missing treatment indicator values.

*(iv) MAR covariate values.* Two MAR mechanisms were considered. Under mechanism MAR1, missing covariate values were simulated with $\Pr(M_{i2} = 1|\mathbf{Z}) = \text{expit}\{-8.2 + 0.8X_{i1}\}(1 - T_i)$. Under mechanism MAR2, $\Pr(M_{i2} = 1|\mathbf{Z}) = \text{expit}\{-13 + 0.8Y_i\}$. These mechanisms set approximately 40% of the subjects' $X_2$ values to missing.

### *4.3.3 Effect estimators*

For all simulated datasets, the Within and Across estimates were obtained as described in the Supplementary Material. Because of their common use, complete case analyses were also performed for comparison. A number of propensity score methods were investigated. The regression estimates of the treatment effect were obtained by linearly regressing the outcome on treatment and the logit of the estimated propensity score. The term matching is used to refer to pair matching performed by selecting for each treated subject a single untreated control without replacement using a greedy nearest neighbour matching algorithm. No restrictions were placed on the maximum acceptable difference between the propensity scores of any two matched subjects. We also performed matching on the logit of the propensity score using a calliper distance of 0.05. A fourth effect estimator was obtained using IPW where treated subjects are weighted by the inverse of their propensity score and untreated subjects by the inverse of its complement. Finally, we applied iterative IPW using a convergence threshold of $10^{-4}$ and a maximum of 100 iterations per dataset (van der Wal, 2011). Calliper matching and iterative IPW were used because matching and IPW are sensitive to practical non-positivity (Cole and Hernán, 2008; van der Wal, 2011).

To improve covariate balance in the presence of practical non-positivity, van der Wal proposed an algorithm in which the dataset is iteratively reweighted (van der Wal, 2011). The idea underpinning this algorithm is as follows. By fitting a propensity model on the weighted dataset, new weights can be estimated that (partially) adjust for the residual confounding. Multiplying these weights with the original yields weights that when applied to the dataset correct for confounding more than the original weights. As the covariate balance improves, the probability of being assigned to treatment becomes less dependent on the covariate values, and so the variance of the log-transformed new weights reduces. The above process is therefore repeated until the variance drops below a convergence threshold.

The iterative inverse probability weighting (IIPW) algorithm was defined in the context of fully observed data. With multiply imputed data, one can apply IIPW within each imputed dataset, in a way consistent with the Within approach, until the algorithm converges within each dataset or until a maximum number of iterations is reached. Alternatively, at each iteration one can average the estimated propensity scores across the imputed datasets, as per the Across approach, before reweighting the imputed datasets. Sample R code for these algorithms is provided in the Supplementary Material.

### *4.3.4 Variance estimation*

An appealing property of the standard multiple imputation approach is that it facilitates estimation of standard errors that reflect both the variability in the data and the uncertainty in the imputations (Rubin, 2004). However, for the Across approach, the between-imputation variance component of Rubin's multiple imputation variance estimator cannot fully capture the uncertainty of the imputations. For example, when there are only missing covariates, the between-imputation variance would be zero, because the same set of propensity scores is used for each dataset.

As an alternative to Rubin's rules for variance estimation, analysts can implement a bootstrapping procedure that is akin to the full mechanism bootstrapping approach described by Efron (Efron and Tibshirani, 1994). Here, bootstrapping is implemented as follows:

1. Sample with replacement $n$ rows from the incomplete dataset $\mathbf{z}$ to obtain a bootstrapped dataset $\mathbf{z}_b$.

2. Impute missing values $m$ times through multiple imputation, producing for $k = 1, 2, ..., m$ an imputed dataset $\mathbf{z}_b^{(k)}$.

3. Apply the analysis procedure (e.g. Within or Across approach) to the $m$ imputed datasets to obtain a single effect estimate $\hat{\beta}_b$ for the bootstrapped dataset.

4. Repeat steps 1–3 $B$ times to obtain $B$ bootstrap replicates.

The bootstrap variance and confidence interval for the effect estimate $\hat{\beta}$ can be obtained from the bootstrap replicates using standard formulae.

For the scenarios with MAR missingness, we estimated variances and confidence intervals using Rubin's rules and the bootstrapping procedure outline above. As discussed, the former can be expected to yield too narrow standard errors and therefore suboptimal coverage. To illustrate this, we applied Rubin's rules for the regression estimators using the modified degrees of freedom formula detailed elsewhere (Barnard and Rubin, 1999; Hughes et al., 2014) to obtain 95% confidence intervals. As for bootstrapping, we calculated bootstrap sample variances and 95% percentile confidence intervals, using the 2.5th and 97.5th percentiles as the lower and upper bounds, based on 1000 bootstrap samples.

### *4.3.5 Performance measures*

The primary performance measure of interest is bias, estimated by the mean deviation of the estimated effect from the true effect of treatment on the outcome ($\beta_{TY}$) across all 1000 simulations, but we also provide empirical variances and mean squared errors (MSE). For the MAR scenarios, coverage probabilities and the mean estimated variances relative to the corresponding empirical variances are also provided. Based on 1000 simulations, the Monte Carlo standard error for the true coverage probability of 0.95 is $\sqrt{}(0.95(1-0.95)/1000) \approx 0.0069$, implying that the estimated coverage probability is expected to lie with 95% probability between 0.936 and 0.964 (Burton et al., 2006). Empirical coverage rates outside this interval provide evidence against the true coverage probabilities being equivalent to the nominal level of 0.95.

## 4.4 Results

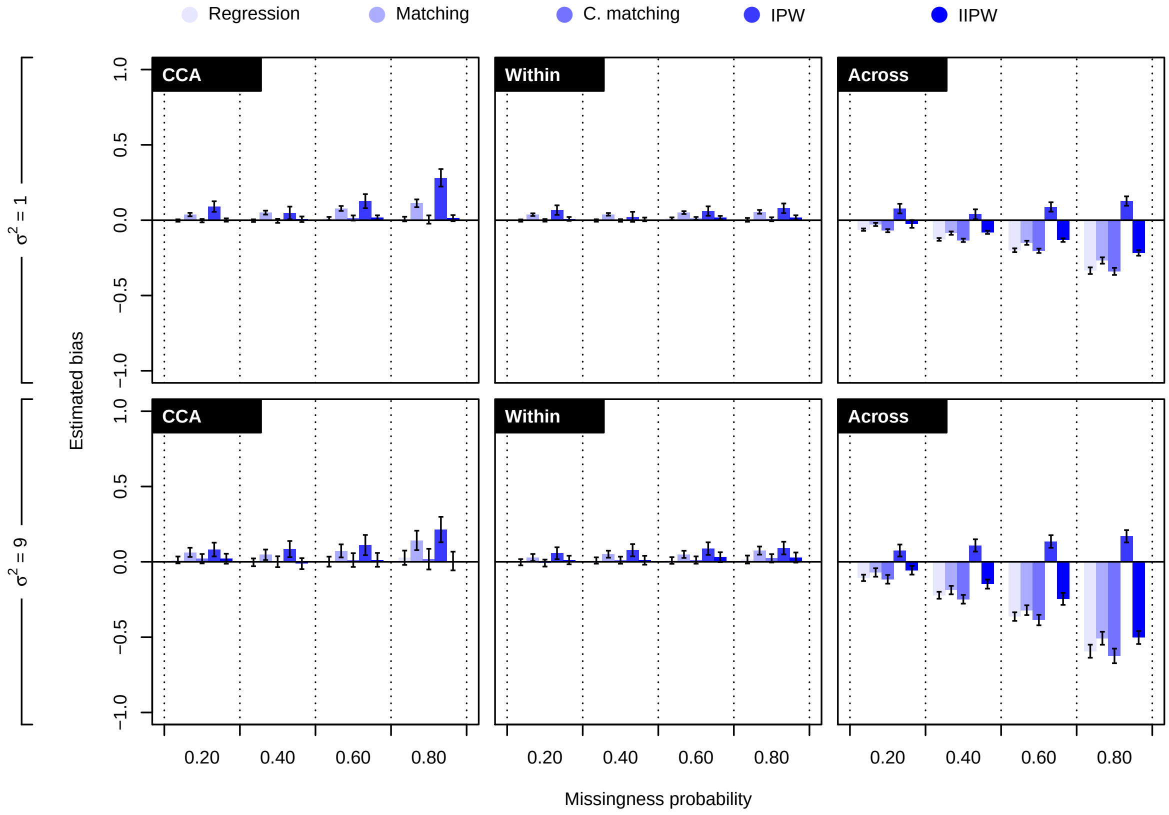
### *4.4.1 Bias*

In this section, we present graphically the estimated biases for the effect estimators of interest. Results on these and other performance measures are presented in tabular form in the Supplementary Material.

#### *1 Missing (MCAR) covariate values*

Figure 4.1 depicts the estimated biases for the scenarios with MCAR covariate data. Apart from those based on matching or IPW, the complete case and Within estimators were not identifiably biased. The Across approach, however, showed substantial bias, especially when either the missingness probability, the residual variance $\sigma^2$ or both were large. The regression-, matching-, calliper matching-, and IIPW-based estimators were all negatively biased for the Across approach. In contrast, Across IPW estimates were on average overestimated. Complete case matching and IPW estimates were also systematically overestimated, with the extent of bias increasing with the extent of missingness.

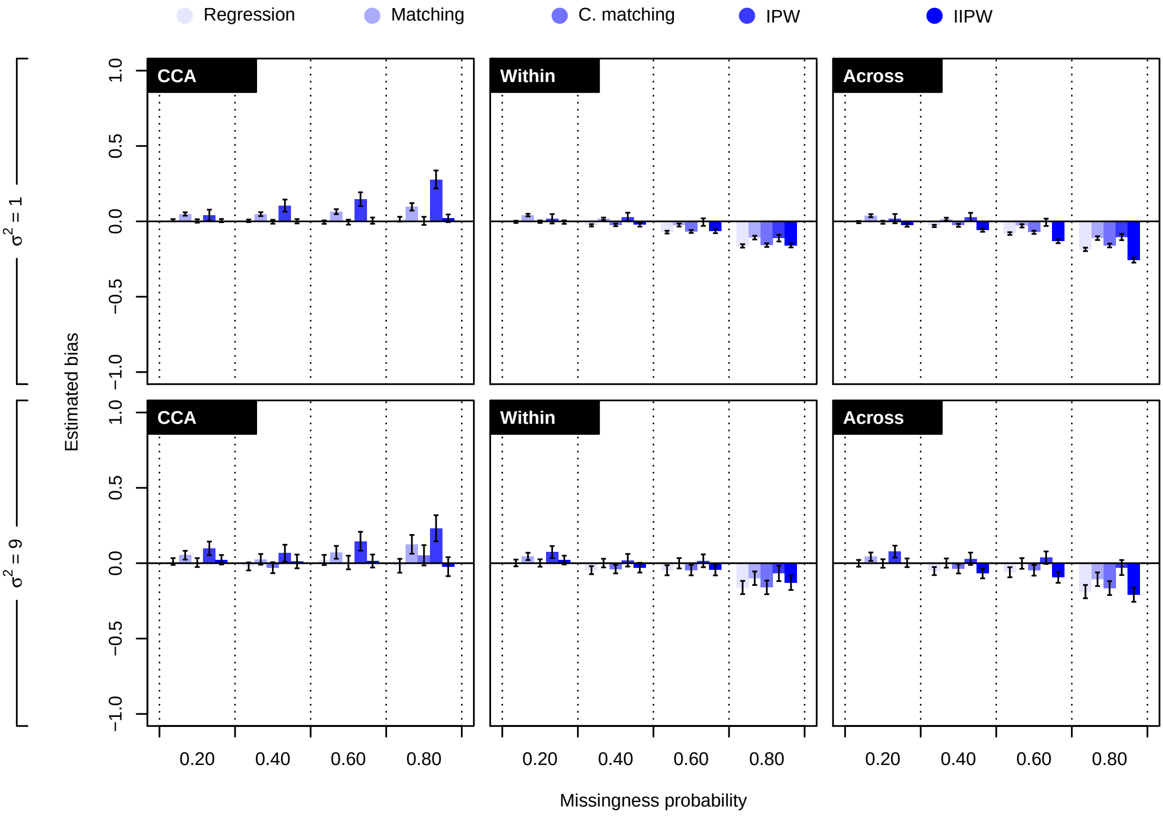
**Figure 4.1:** Biases of treatment effect estimators for various degrees of missing (MCAR) covariate data and residual variances $\sigma^2$. *Abbreviations:* C. matching, calliper matching; IPW, inverse probability weighting; IIPW, iterative inverse probability weighting; CCA, complete case analysis. Error bars represent 95% confidence intervals for the simulation estimates of bias.



## *2 Missing (MCAR) treatment indicator values*

Figure 4.2 depicts the estimated biases for the scenarios with MCAR treatment indicator values. The Across and Within estimates were on average highly similar. Apparent from the figure is also the trend that as the percentage of incomplete cases increases, the treatment effect becomes on average progressively more underestimated by both the Across and Within estimators. Conversely, the complete case matching and IPW estimators systematically overestimated the

**Figure 4.2:** Biases of treatment effect estimators for various degrees of missing (MCAR) treatment indicator values and residual variances $\sigma^2$. *Abbreviations:* C. matching, calliper matching; IPW, inverse probability weighting; IIPW, iterative inverse probability weighting; CCA, complete case analysis. Error bars represent 95% confidence intervals for the simulation estimates of bias.



treatment effect, particularly for large missingness probabilities.

*3 Missing (MCAR) outcome values*

Figure 4.3 depicts the estimated biases for the scenarios with MCAR outcomes. For all propensity score methods, the Across and Within estimators yielded identical results. Again, the complete case matching and IPW estimators showed bias, particularly when the extent of missingness was large. The corresponding Within and Across estimators were less biased. The regression-, calliper matching-, and IIPW-based estimators resulted in minimal bias.

*4 Missing (MAR) covariate values*

Figure 4.4 depicts the estimated biases for the scenarios with MAR covariate data. The complete case matching and IPW estimators generally showed more bias than in the corresponding MCAR covariate settings with a comparable proportion of incomplete records (40%). The regression-, calliper matching-, and IIPW-based estimators showed minimal bias for both the complete case analysis and Within approach when the missingness of $X_2$ depended on $X_1$ and $T$ (mechanism MAR1). As for the scenarios where the missingness dependend on the outcome $Y$ (MAR2), the Within but not the complete case approach yielded estimates close to the true treatment effect. As before, Across estimates were systematically too low for the regression-, matching-, calliper matching-, and IIPW-based estimators.

### *4.4.2 Other performance measures*

In general, the Within estimators were associated with the smallest empirical variances and MSEs. The simulations also illustrate the implications of using Rubin's rules in estimating the variance. The variances of the Across regression estimators were underestimated and the coverage probability too low (see Supplementary Material). Conversely, when applying the bootstrapping procedure, the estimated variances were on average close to the respective empirical variances. Despite generally adequate coverage probabilities for the Within approach, the variances for calliper matching-, and IIPW-based estimators were on average overestimated.

## 4.5 Discussion

Our primary focus was on examining the relative performance of two approaches to implementing propensity score methods following multiple imputation.

**Figure 4.3:** Biases of treatment effect estimators for various degrees of missing (MCAR) outcomes and residual variances $\sigma^2$. *Abbreviations:* C. matching, calliper matching; IPW, inverse probability weighting; IIPW, iterative inverse probability weighting; CCA, complete case analysis. Error bars represent 95% confidence intervals for the simulation estimates of bias.

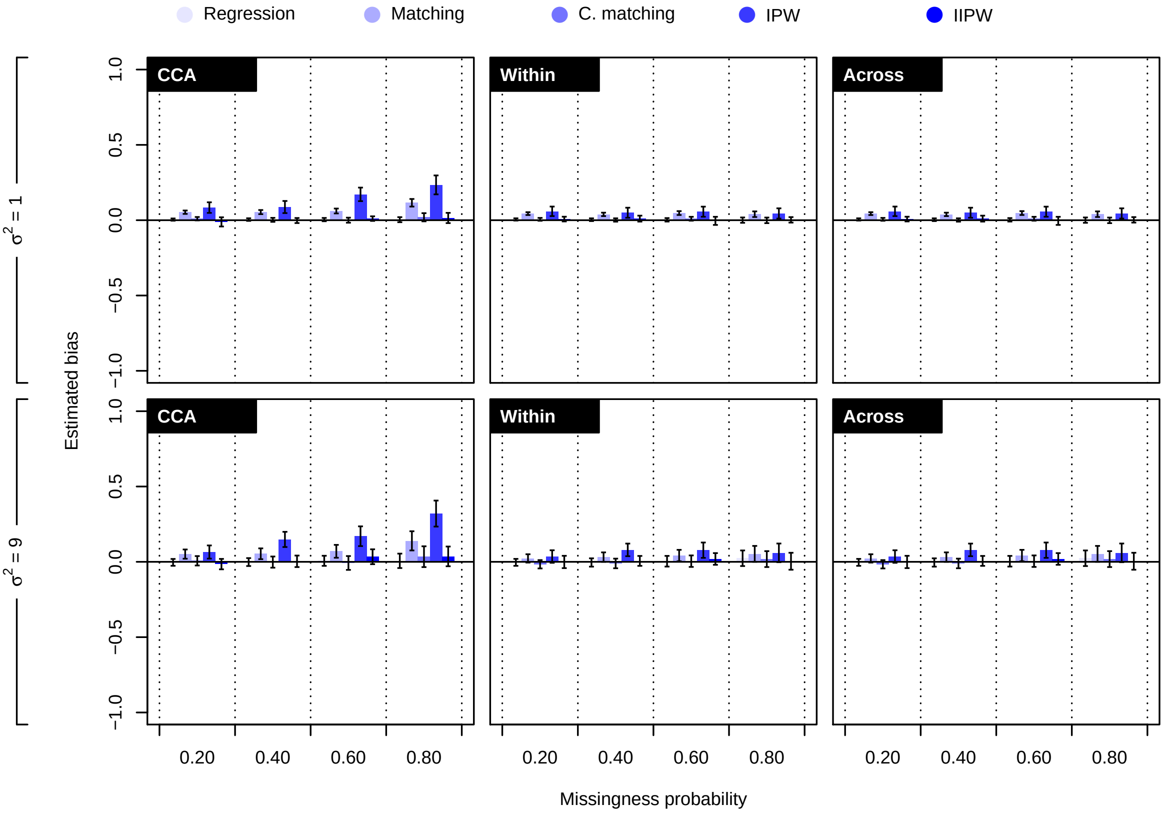
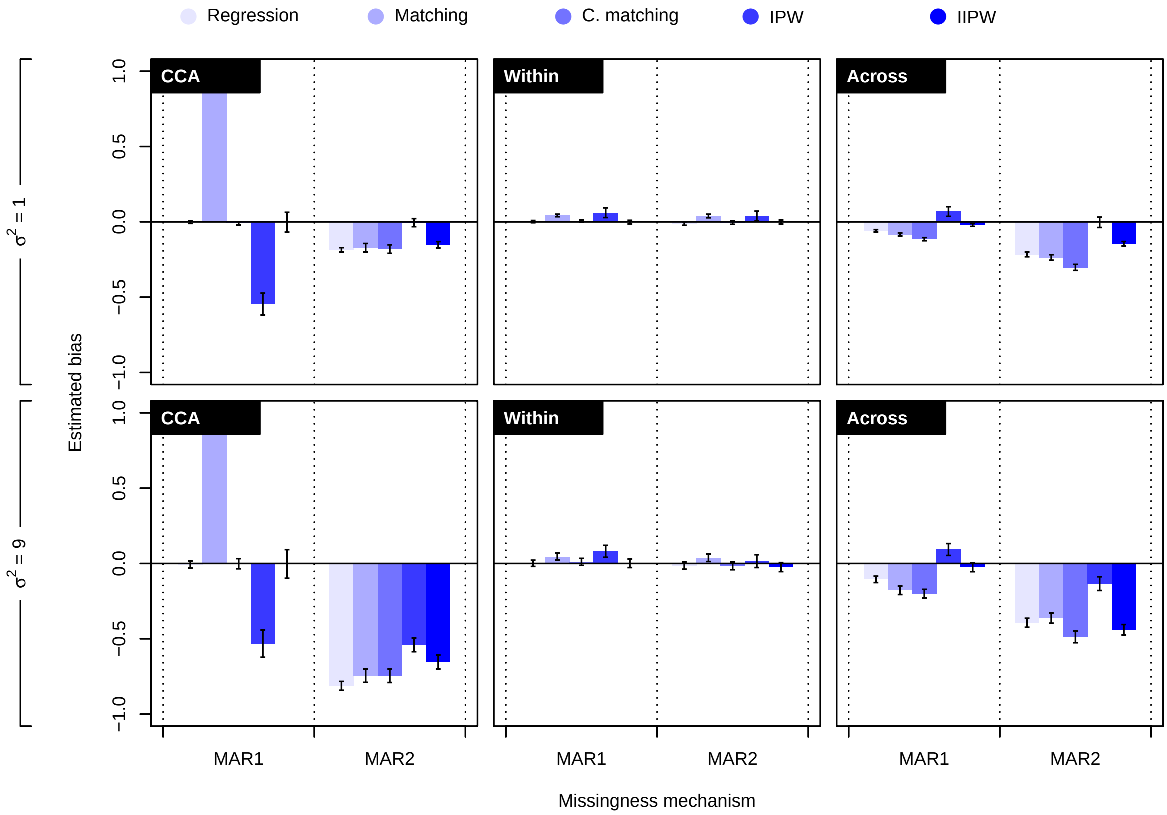


**Figure 4.4:** Biases of treatment effect estimators under various (MAR) missingness mechanisms and residual variances $\sigma^2$. *Abbreviations:* C. matching, calliper matching; IPW, inverse probability weighting; IIPW, iterative inverse probability weighting; CCA, complete case analysis. Under mechanism MAR1, the missingness of $X_2$ depends on $X_1$ and $T$ only. Under MAR2, the missingness depends on $Y$ only. Both MAR1 and MAR2 result in ~40% incomplete records. Error bars represent 95% confidence intervals for the simulation estimates of bias.



Although the Across approach has been applied in practice, our simulations show that it fails in settings with missing confounder data, even when the missingness is completely at random and complete case estimators are unbiased.

As described in the Supplementary Material, untreated subjects with propensity scores that are by random variability underestimated are more likely to be selected as matches, or are assigned greater weight in IIPW, than subjects whose propensity score is overestimated. This problem of random variability is inherent to semi-parametric propensity score methods, and is not expected to introduce bias when regression adjustment is used. However, its impact was negligible in our simulations, because the calliper matching and IIPW estimates were highly similar to the regression estimates. The second explanation for the discrepancy in bias between the approaches rests on the resemblance of the Across approach to conditional mean imputation in the context of missing covariate data. This explanation is consistent with our observations that the Across approach showed more bias for larger missingness probabilities and larger residual variances.

Conversely, in the absence of missing confounder data, the Across approach is not comparable to conditional mean imputation. Instead, the bias observed in settings with missing treatment indicator values probably is largely attributable to a phenomenon, known as separation or 'perfect prediction', that is associated with regression models for categorical responses. Separation occurs if the responses, here treatment status, can be perfectly separated by a single predictor or a linear combination of predictors. The problem lies with the Normal approximation to the posterior distribution of the parameters of the logistic regression model that is used by the software to predict missing treatment indicator values. When in the presence of separation, logistic regression is applied to the complete cases, modelling the probability of being assigned to treatment as $\Pr(T_i = 1|X_{i1}, X_{i2}, Y_i) = \text{expit}\{\alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 Y_i\}$, then we can find an infinite sequence of parameter specifications with monotonically increasing likelihood converging to unity, such that for at least one parameter $\alpha$ the estimate $\hat{\alpha}$ tends to infinity (Albert and Anderson, 1984; Heinze and Schemper, 2002). Hence, the maximum likelihood estimate does not exist. Nevertheless, given the near-flat nature of the likelihood, typically very large values for the maximum likelihood estimate $\hat{\alpha}$ and its variance are returned by standard software. If the Normal approximation to the posterior distribution of the parameters is applied, then it is not unlikely that values are drawn such that in the imputation step subjects with incomplete data are assigned to the treatment group whilst the observed data clearly suggests that these subjects should be assigned to the control group (White et al., 2010). In other words, the Normal approximation to the posterior distribution is poor. One way to prevent these implausible imputations is to add to the dataset a few observations such that separation is no longer present and with such small weights that the impact on the imputation model is limited (White et al., 2010). `mice` implements such a data augmentation method to deal with this phenomenon (van Buuren and Groothuis-Oudshoorn, 2011; van Buuren, 2012), but we suspect that in our simulations the impact of the weights was large enough to produce bias.

MAR1 is an example of a mechanism that accentuates practical non-positivity. Under this mechanism, untreated subjects with large $X_1$ values are more likely to be assigned missing $X_2$ values than others. When untreated subjects have systematically lower $X_2$ values even before the introduction of missingness, the consequence of this mechanism is that the propensity score distributions of groups of treated and untreated subjects become more distinct. As a result, estimators that are sensitive to practical non-positivity (e.g. matching and IPW) become more biased. Note that the matching and IPW methods described in Section 4.3.3 are estimators of the average effect on the treated (ATT) and the average effect (ATE) on all subjects, respectively (Williamson et al., 2012). A sufficient condition for these measures of effect to coincide is that of collapsibility and treatment effect homogeneity. This joint condition is met in our simulations. The assumptions of ATT and ATE estimators with respect to positivity are, however, not the same. ATE estimators require the covariate distributions of the treated and untreated to have common support, whereas ATT estimators require only the support of the treated to be shared by that of the untreated but not vice versa (Lechner, 2008) This largely explains the discrepancies in bias across PS methods in our simulations.

Daniel et al. (2012) show how causal diagrams can be used to infer that in nearly all scenarios considered here conditioning on the complete cases (i.e. prior to PS matching, IPW, or IIPW) does not itself induce bias. It follows that when other sources of bias, here practical non-positivity and confounding, are adequately addressed, the treatment effect can be validly estimated. Among the missing data mechanisms considered, it is only MAR2 that biases complete case analyses, namely by inducing a relation between treatment status and the (unmeasured) error on the outcome through collider stratification.

Although unbiasedness is arguably more crucial than valid variance estimation, sufficiently large variances, even if they can be estimated validly, may render unbiased estimators of little practical use (Burton et al., 2006). In our simulations, the Within approach was superior or comparable to the Across in terms of either criterion. Another drawback of the latter approach is the difficulty in making inferences as to the precision of effect estimates.

Bootstrapping may provide correct standard errors, but we acknowledge that this approach is computationally intensive. Further, because coverage is affected by the bias in both variance and effect estimation, it is likely to be poor in general for Across estimators. Bootstrapping for the (calliper) matching estimators here yielded slightly overestimated variances and conservative empirical coverage rates. A similar phenomenon was observed by Austin and Small (2014). The bootstrapping procedure defined in Section 4.3.4 resembles the 'complex bootstrap' of Austin and Small. The rather large discrepancies between the mean estimated variances and the empirical variances for the IIPW estimators are possibly attributable to the unstable nature of inverse probability weighting. Further investigation and development of bootstrapping approaches to variance estimation for the (calliper) matching and (iterative) IPW estimators represent an interesting direction for future research.

Our findings contrast with those of Mitra and Reiter (2016). A crucial difference between the simulations is the inclusion of the outcome in the model used to impute missing covariate values. Failing to include the outcome leads to imputed datasets that do not reflect the association between covariate and outcome that would have been observed had there been no missing values. The consequence of this is that if one adjusts for the imputed covariate values to estimate the treatment effect, the variation in outcomes between the treatment groups that is due to the partially unobserved covariate would in part be attributed to the differences in treatment status.

As with any simulation study, an important limitation of this study is the potentially limited generalisability. The scenarios considered here represent only a small and simplified subset of those likely to be encountered in applied research. Some of the missingness probabilities that were studied are probably unrealistic, and only a single sample size was considered. Practical non-positivity and separation are perhaps less relevant in settings with larger samples and fewer incomplete cases. Furthermore, we considered only two covariates and assumed that the propensity score and imputation models could be correctly specified. Applied researchers do not have the luxury of knowing the data generating and missing data mechanisms and often need to assess and account for multiple sources of bias. However, rather than scrutinising methods for these issues in isolation only, it may be interesting to additionally study how they perform in combination. Conducting simulations for specific scenarios of interest may be particularly desirable given the limited generalisability of our results. If these are not possible, we advise researchers not to use the Across approach as the default method, because it appears to offer no advantage over the Within method.

## 4.6 Conclusion

In medical research, confounding and missing data are common problems that often occur simultaneously. When multiple imputation is to be followed by the implementation of a propensity score method, researchers could apply the Across and Within approaches. The present study highlights a number of aspects of the Across approach that render it suboptimal. Our simulations indicate that the Within approach is superior to the Across approach in terms of both bias and variance in settings with missing confounder data. For incomplete treatment and/or outcome data, the approaches yield similar estimates. We advise researchers not to use the Across approach as the default method, because even in MCAR settings, this may yield biased effect estimates. Finally, when matching or IPW are the propensity methods of choice, we recommend practical non-positivity to be adequately addressed, e.g. by using a narrow calliper or an iterative reweighting algorithm.

## References

Albert, A. and J. Anderson (1984): "On the existence of maximum likelihood estimates in logistic regression models," *Biometrika*, 71, 1–10.

Austin, P. C. (2011): "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate behavioral research*, 46, 399–424.

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

Barnard, J. and D. B. Rubin (1999): "Miscellanea. small-sample degrees of freedom with multiple imputation," *Biometrika*, 86, 948–955.

Brown, J. B., C. M. Leeper, J. L. Sperry, A. B. Peitzman, T. R. Billiar, B. A. Gaines, and M. L. Gestring (2016): "Helicopters and injured kids: Improved survival with scene air medical transport in the pediatric trauma population," *Journal of trauma and acute care surgery*, 80, 702–710.

Burton, A., D. G. Altman, P. Royston, and R. L. Holder (2006): "The design of simulation studies in medical statistics," *Statistics in medicine*, 25, 4279–4292.

Chiu, P., J. M. Schaffer, P. E. Oyer, M. Pham, D. Banerjee, Y. J. Woo, and R. Ha (2016): "Influence of durable mechanical circulatory support and allosensitization on mortality after heart transplantation," *The Journal of Heart and Lung Transplantation*, 35.

Cole, S. R. and M. A. Hernán (2008): "Constructing inverse probability weights for marginal structural models," *American journal of epidemiology*, 168, 656–664.

Daniel, R. M., M. G. Kenward, S. N. Cousens, and B. L. De Stavola (2012): "Using causal diagrams to guide analysis in missing data problems," *Statistical methods in medical research*, 21, 243–256.

Efron, B. and R. J. Tibshirani (1994): *An introduction to the bootstrap*, CRC press.

Ekström, N., A.-M. Svensson, M. Miftaraj, S. Franzén, B. Zethelius, B. Eliasson, and S. Gudbjörnsdottir (2016): "Cardiovascular safety of glucose-lowering agents as add-on medication to metformin treatment in type 2 diabetes: report from the swedish national diabetes register," *Diabetes, Obesity and Metabolism*, 18, 990–998.

Gregory, E. F., S. M. Gross, T. Q. Nguyen, A. M. Butz, and S. B. Johnson (2016): "Wic participation and breastfeeding at 3 months postpartum," *Maternal and child health journal*, 20, 1–10.

Heinze, G. and M. Schemper (2002): "A solution to the problem of separation in logistic regression," *Statistics in medicine*, 21, 2409–2419.

Hughes, R., J. Sterne, and K. Tilling (2014): "Comparison of imputation variance estimators," *Statistical methods in medical research*, 25, 2541–2557.

Kutney-Lee, A., G. Melendez-Torres, M. D. McHugh, and B. M. Wall (2014): "Distinct enough? a national examination of catholic hospital affiliation and patient perceptions of care," *Health care management review*, 39, 134.

Lechner, M. (2008): "A note on the common support problem in applied evaluation studies," *Annales d'Économie et de Statistique*, 91/92, 217–235.

Leyrat, C., S. R. Seaman, I. R. White, I. Douglas, L. Smeeth, J. Kim, M. Resche-Rigon, J. R. Carpenter, and E. J. Williamson (2017): "Propensity score analysis with partially observed covariates: How should multiple imputation be used?" *Statistical Methods in Medical Research*, 0, 1–17.

Mitra, R. and J. P. Reiter (2016): "A comparison of two methods of estimating propensity scores after multiple imputation," *Statistical methods in medical research*, 25, 188–204.

Moons, K. G. M., A. R. T. Donders, T. Stijnen, and F. Harrell (2006): "Using the outcome for imputation of missing predictor values was preferred." *Journal of clinical epidemiology*, 59, 1092–101.

Neuderth, S., B. Schwarz, C. Gerlich, M. Schuler, M. Markus, and M. Bethge (2016): "Work-related medical rehabilitation in patients with musculoskeletal disorders: the protocol of a propensity score matched effectiveness study (eva-wmr, drks00009780)," *BMC Public Health*, 16, 804.

Olszewski, A. J., R. Shrestha, and J. J. Castillo (2015): "Treatment selection and outcomes in early-stage classical hodgkin lymphoma: Analysis of the national cancer data base," *Journal of Clinical Oncology*, 33, 625–633.

Penning de Vries, B. and R. Groenwold (2016): "Comments on propensity score matching following multiple imputation," *Statistical methods in medical research*, 25, 3066–3068.

R Core Team (2016): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `https://www.R-project.org/`.

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

Rubin, D. B. (2004): *Multiple imputation for nonresponse in surveys*, volume 81, John Wiley & Sons.

Schafer, J. L. and J. W. Graham (2002): "Missing data: our view of the state of the art." *Psychological methods*, 7, 147.

Sterne, J. A., I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter (2009): "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls," *BMJ*, 338, b2393.

Stürmer, T., M. Joshi, R. J. Glynn, J. Avorn, K. J. Rothman, and S. Schneeweiss (2006): "A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods," *Journal of clinical epidemiology*, 59, 437–e1.

Sulkowski, J. P., J. N. Cooper, A. Congeni, E. G. Pearson, B. C. Nwomeh, E. J. Doolin, M. L. Blakely, P. C. Minneci, and K. J. Deans (2014): "Single-stage versus multi-stage pull-through for hirschsprung's disease: Practice trends and outcomes in infants," *Journal of pediatric surgery*, 49, 1619–1625.

Sulkowski, J. P., J. N. Cooper, E. M. Duggan, O. Balci, S. P. Anandalwar, M. L. Blakely, K. Heiss, S. Rangel, P. C. Minneci, and K. J. Deans (2015): "Does timing of neonatal inguinal hernia repair affect outcomes?" *Journal of pediatric surgery*, 50, 171–176.

van Buuren, S. (2012): *Flexible imputation of missing data*, CRC Press.

van Buuren, S. and K. Groothuis-Oudshoorn (2011): "mice: Multivariate imputation by chained equations in r," *Journal of Statistical Software*, 45, 1–67, URL `http://www.jstatsoft.org/v45/i03/`.

van der Wal, W. M. (2011): *Causal modeling in epidemiological practice*, Phd thesis, University of Amsterdam.

White, I. R., R. Daniel, and P. Royston (2010): "Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables," *Computational statistics & data analysis*, 54, 2267–2275.

Williamson, E., R. Morley, A. Lucas, and J. Carpenter (2012): "Propensity scores: from naive enthusiasm to intuitive understanding," *Statistical methods in medical research*, 21, 273–293.

## Supplementary Material

This Supplementary Material has three parts. The first contains a discussion of the Within and Across approaches; the second part provides sample R code for the iterative inverse probability weighting estimators for the complete case, Across and Within approaches; and the third part provides the results on all performance measures.

## S4.1 Within and Across approaches

Two approaches to implementing propensity score methods in the presence of missing data are described: the Within and Across approaches (Mitra and Reiter, 2016). The first step in the analysis procedure is to estimate within each completed dataset a vector of propensity scores $\boldsymbol{e}^{(k)} = (e_1^{(k)}, e_2^{(k)}, ..., e_n^{(k)})$ typically using logistic regression. In the Within approach, any propensity score method is implemented within each completed dataset using $\boldsymbol{e}^{(k)}$, yielding $\hat{\beta}_W^{(k)}$. The Within estimate $\hat{\beta}_W$ is defined as the average of $\hat{\beta}_W^{(1)}, \hat{\beta}_W^{(2)}, ..., \hat{\beta}_W^{(m)}$. In the Across approach, the propensity score method is implemented within each completed dataset, now using $\boldsymbol{e}^A = (e_1^A, e_2^A, ..., e_n^A)$, where $e_i^A = \sum_{k=1}^m e_i^{(k)}/m$. As in the Within approach, the resulting estimates $\hat{\beta}_A^{(1)}, \hat{\beta}_A^{(2)}, ..., \hat{\beta}_A^{(m)}$ are averaged, yielding the single estimate $\hat{\beta}_A$. This procedure deviates slightly from the Across approach described by Mitra and Reiter (2016). The modified Across approach described here is equivalent to the original procedure when $T$ and $Y$ are fully observed, but can additionally accommodate missings on $T$ and/or $Y$. Henceforth, we will refer to this modified procedure simply as the Across approach.

Note that only the Within approach fully adheres to Rubin's original MI algorithm, where averaging across imputations is deferred until the last step. In the context of semi-parametric propensity score methods, such as matching, this may seem unsatisfactory. Untreated subjects who would be considered unsuitable matches based on their 'true' propensity scores, may be included in the matched set because by random variability their estimated propensity scores, based on the imputed data, better resemble the treated subjects' propensity scores. When the outcome is associated with the propensity score, this may then lead to bias. Intuitively, the Across approach may then be preferable because of the lesser reliance on random variability. However, for large $m$ the Across approach is comparable to conditional mean imputation, which as we now illustrate may introduce bias.

Consider for simplicity the case with only a single continuous covariate $X$, and suppose that the treatment-outcome effect can be parameterised in the linear regression

$$\mathbb{E}(Y|T,X) = \beta_0 + \beta_1 T + \beta_2 X$$

where $\beta_1$ is the parameter of interest. Further, assume that the relationship between the probability of being assigned to treatment ($T = 1$) given $X$ can be modelled by a logistic function

$$\Pr(T = 1|X) = \frac{\exp(\alpha_0 + \alpha_1 X)}{1 + \exp(\alpha_0 + \alpha_1 X)}$$

In this case, we may rewrite the treatment-outcome model in terms of a linear transformation logit $e(X) = \alpha_0 + \alpha_1 X$ of the covariate values, the logit of the propensity score $e(X)$,

$$\mathbb{E}(Y|T,X) = (\beta_0 - \beta_1\alpha_1^{-1}\alpha_0) + \beta_1 T + \beta_1\alpha_1^{-1}\text{logit}\{e(X)\}$$

The ordinary least squares estimator of the true treatment effect $\beta_1$ is unbiased if the regressors of the linear model are $T$ and $X$, or $T$ and logit $e(X)$. A similar observation can be made in the case of multiple covariates (Wan and Mitra, 2016). However, if we impute missing $X$ values using conditional mean imputation, and regress $Y$ on $T$ and the (linearly transformed) imputed covariates to estimate the treatment effect, then, as illustrated in Figure S4.1, the estimator will be biased—provided that the conditional variance of $Y$ given $T$ and $X$ is greater than zero and treatment assignment depends on $X$.

Likewise, in the case of missing (e.g. MCAR) $X$ values, averaging (transformed) imputed values across many multiply imputed datasets (i.e. effectively conditional mean imputation) also renders the effect estimator biased. The crux of the matter lies in that the default imputation model is the linear regression with $X$ as the dependent variable and $T$ and $Y$ as the independent variables, whereas the analysis model regards $Y$ as the dependent variable. Switching dependent and independent variables results in best fit equations that are not in general equivalent (unless orthogonal regression is used). Bias can therefore also be expected for the Across approach, because in the context of missing covariate data it is comparable to conditional mean imputation, except that taking the logit of the average propensity score is not the same as taking the average of the logit of propensity scores.

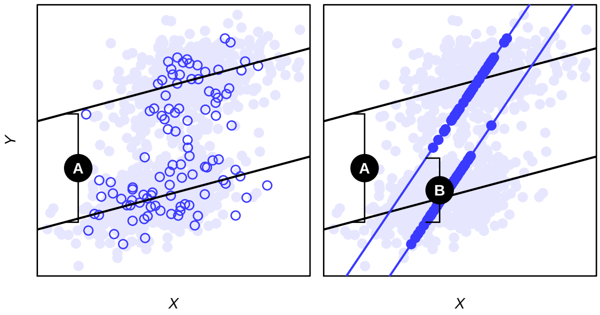
It follows from this discussion that the Across approach should be similar in terms of bias to the Within approach when there are only missings on $T$ or $Y$, because averaging propensity scores across datasets does not affect the variability in the imputed values of these variables. Moreover, if there were missing outcome values only, the imputation model with $Y$ as the outcome and $T$ and $X$ as the regressors would clearly be compatible with the analysis model. Hence, even conditional mean imputation of missing $Y$ would suffice for unbiased estimation of the treatment effect, provided the missingness is ‘ignorable’. Furthermore, when treatment and covariates are fully observed, clearly the propensity scores do not differ across imputed datasets, and so the Across and Within estimators yield identical effect estimates.

## S4.2 Iterative inverse probability weighting: R code

```
truncate <- function(x,left=0.05,right=0.95){
  Q <- quantile(x,probs=c(left,right))
  x[x>Q[2]] <- Q[2]; x[x<Q[1]] <- Q[1]
  return(x)
}

IIPTW <- function(
  # Iterative Inverse Probability of Treatment Weighting
  # Returns a list with weights and number of iterations
  data, # a dataframe containing columns T, X1, X2 and Y
  formula =‘T˜X1+X2’,
  T =‘T’
  left =0,
  right =1,
  cstop =1e-4,
```

**Figure S4.1:** A single random sample (left) with missing completely at random (MCAR) covariate data imputed using conditional mean imputation (right). A valid treatment effect (A) is obtained by the regression of $Y$ on $T$ and $X$ (the analysis model) applied to the complete cases. Applying the analysis model to the subset with covariate values imputed through conditional mean imputation yields a biased treatment effect estimate (B).



```
  maxit =100,
  estimand =‘ATE’
  ){
  warning <- FALSE
  it <- 1
  n <- nrow(data)
  w <- rep(1,n)
  for(i in 1:maxit){
    psnew <- predict(glm(formula=as.formula(formula),family=binomial,data=
        data,weights=w),type=‘response’)
    if(estimand==‘ATE’){wnew <- ifelse(data[,T]==1,1/psnew,1/(1-psnew))}
    if(estimand==‘ATT’){wnew <- ifelse(data[,T]==1,1,psnew/(1-psnew))}
    if(estimand==‘ATU’){wnew <- ifelse(data[,T]==1,(1-psnew)/psnew,1)}
    if(i>=2 && var(log(wnew))<=cstop){it <- i; break}
    if(i==maxit){it <- i; warning <- TRUE}
    w <- truncate(w*wnew,left,right)
    w <- w/mean(w)
  }
  if(warning==TRUE){warning(‘Algorithm did not converge’);it<-NA}
  return(list(w=data$w,it=it))
}

IIPTW.A <- function(
  # Iterative Inverse Probability of Treatment Weighting
  # consistent with the Across approach
  # Returns a list with weights for each imputed dataset and the number of
      iterations
  data, # a list of multiply imputed datasets
  formula =‘T~X1+X2’,
  T =‘T’,
  left =0,
  right =1,
  cstop =1e-4,
  maxit =100,
  estimand =‘ATE’
  ){
  warning <- FALSE
  it <- 1
  m <- length(data)
  n <- nrow(data[[1]])
  for(i in 1:m){data[[i]]$w <- rep(1,n)}
  psnew <- matrix(nrow=n,ncol=m)
  for(i in 1:maxit){
    for(u in 1:m){
      psnew[,u] <-  predict(glm(formula=as.formula(formula),family=binomial
          ,data=data[[u]],weights=data[[u]]$w),type=‘response’)
    }
    psnewA <- apply(psnew,1,mean)
    for(u in 1:m){
      if(estimand==‘ATE’){data[[u]]$wnew <- ifelse(data[[u]][,T]==1,1/
          psnewA,1/(1-psnewA))}
      if(estimand==‘ATT’){data[[u]]$wnew <- ifelse(data[[u]][,T]==1,1,
          psnewA/(1-psnewA))}
      if(estimand==‘ATU’){data[[u]]$wnew <- ifelse(data[[u]][,T]==1,(1-
          psnewA)/psnewA,1)}
    }
    if(i>=2 && prod(unlist(lapply(data,FUN=function(x){var(log(x$wnew))<=
        cstop})))==1){it <- i; break}
    if(i==maxit){it <- i; warning <- TRUE}
    for(u in 1:m){
      data[[u]]$w <- truncate(data[[u]]$w*data[[u]]$wnew,left=left,right=
          right)
      data[[u]]$w <- data[[u]]$w/mean(data[[u]]$w)
    }
  }
  if(warning==TRUE){warning(‘Algorithm did not converge.’);it<-NA}
  return(list(w=lapply(data,function(x)x$w),it=it))
}

IIPTW.W <- function(
  # Iterative Inverse Probability of Treatment Weighting
  # consistent with the Within approach
  # Returns a list with weights for each imputed dataset and the number of
      iterations
  data, # a list of multiply imputed datasets
  formula =‘T~X1+X2’,
  T =‘T’,
  left =0,
  right =1,
  cstop =1e-4,
  maxit =100,
  estimand =‘ATE’
  ){
  out <- lapply(data,FUN=IIPTW,formula=formula,T=T,left=left,right=right,
      cstop=cstop,maxit=maxit,estimand=estimand)
  w <- lapply(out,function(x)x$w)
  itmax <- max(unlist(lapply(out,function(x)x$it)))
  return(list(w=w,it=itmax))
}
```

## S4.3 Results

**Table S4.1:** Performance of treatment effect estimators for various degrees $p$ of missing (MCAR) covariate data and residual variances $\sigma^2$. *Abbreviations:* CCA, complete case analysis; $p$, missingness probability (%); PS method, propensity score method; $\bar{\hat{\beta}} - \beta$, estimated bias; Emp. var., empirical variance; MSE, empirical mean squared error; C. matching, calliper matching; IPW, inverse probability weighting; IIPW, iterative inverse probability weighting.

| | | | CCA | | | Within | | | Across | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma^2$ | $p$ | PS method | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE |
| 1 | 20 | Regression | -0.002 | 0.016 | 0.016 | -0.003 | 0.015 | 0.015 | -0.063 | 0.016 | 0.020 |
| | | Matching | 0.037 | 0.033 | 0.034 | 0.037 | 0.021 | 0.022 | -0.028 | 0.029 | 0.030 |
| | | C. matching | -0.003 | 0.032 | 0.032 | -0.001 | 0.019 | 0.019 | -0.069 | 0.028 | 0.033 |
| | | IPW | 0.091 | 0.322 | 0.330 | 0.067 | 0.263 | 0.268 | 0.077 | 0.262 | 0.268 |
| | | IIPW | 0.001 | 0.031 | 0.031 | 0.008 | 0.043 | 0.043 | -0.025 | 0.172 | 0.173 |
| | 40 | Regression | -0.003 | 0.020 | 0.020 | -0.002 | 0.017 | 0.017 | -0.128 | 0.022 | 0.039 |
| | | Matching | 0.050 | 0.044 | 0.046 | 0.039 | 0.020 | 0.021 | -0.086 | 0.032 | 0.040 |
| | | C. matching | -0.005 | 0.046 | 0.046 | -0.002 | 0.019 | 0.019 | -0.134 | 0.032 | 0.050 |
| | | IPW | 0.048 | 0.461 | 0.463 | 0.022 | 0.288 | 0.288 | 0.039 | 0.288 | 0.289 |
| | | IIPW | 0.006 | 0.088 | 0.088 | 0.005 | 0.039 | 0.039 | -0.081 | 0.029 | 0.036 |
| | 60 | Regression | 0.011 | 0.029 | 0.030 | 0.010 | 0.021 | 0.021 | -0.199 | 0.042 | 0.082 |
| | | Matching | 0.078 | 0.071 | 0.077 | 0.051 | 0.024 | 0.026 | -0.150 | 0.048 | 0.071 |
| | | C. matching | 0.015 | 0.075 | 0.075 | 0.012 | 0.024 | 0.024 | -0.203 | 0.054 | 0.096 |
| | | IPW | 0.127 | 0.566 | 0.582 | 0.061 | 0.257 | 0.261 | 0.088 | 0.254 | 0.262 |
| | | IIPW | 0.017 | 0.064 | 0.064 | 0.017 | 0.036 | 0.036 | -0.132 | 0.038 | 0.056 |

Table S4.1 continued.

| | | | CCA | | | Within | | | Across | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma^2$ | $p$ | PS method | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE |
| 1 | 80 | Regression | 0.008 | 0.060 | 0.060 | 0.002 | 0.042 | 0.042 | -0.335 | 0.128 | 0.240 |
| | | Matching | 0.112 | 0.172 | 0.185 | 0.055 | 0.040 | 0.043 | -0.268 | 0.117 | 0.189 |
| | | C. matching | 0.004 | 0.198 | 0.198 | 0.006 | 0.045 | 0.045 | -0.340 | 0.143 | 0.258 |
| | | IPW | 0.281 | 0.884 | 0.963 | 0.079 | 0.264 | 0.271 | 0.127 | 0.254 | 0.270 |
| | | IIPW | 0.014 | 0.107 | 0.108 | 0.018 | 0.053 | 0.054 | -0.217 | 0.089 | 0.136 |
| 9 | 20 | Regression | 0.013 | 0.134 | 0.134 | -0.002 | 0.112 | 0.112 | -0.106 | 0.116 | 0.127 |
| | | Matching | 0.063 | 0.242 | 0.246 | 0.029 | 0.142 | 0.142 | -0.071 | 0.209 | 0.214 |
| | | C. matching | 0.021 | 0.251 | 0.251 | -0.008 | 0.139 | 0.139 | -0.116 | 0.217 | 0.231 |
| | | IPW | 0.082 | 0.539 | 0.545 | 0.057 | 0.405 | 0.409 | 0.075 | 0.410 | 0.416 |
| | | IIPW | 0.021 | 0.284 | 0.285 | 0.012 | 0.207 | 0.207 | -0.056 | 0.220 | 0.223 |
| | 40 | Regression | -0.003 | 0.174 | 0.174 | 0.009 | 0.115 | 0.115 | -0.222 | 0.139 | 0.188 |
| | | Matching | 0.047 | 0.314 | 0.316 | 0.051 | 0.140 | 0.142 | -0.188 | 0.202 | 0.237 |
| | | C. matching | 0.001 | 0.339 | 0.339 | 0.011 | 0.143 | 0.143 | -0.249 | 0.217 | 0.279 |
| | | IPW | 0.085 | 0.747 | 0.754 | 0.077 | 0.430 | 0.436 | 0.109 | 0.433 | 0.445 |
| | | IIPW | -0.012 | 0.346 | 0.346 | 0.010 | 0.224 | 0.224 | -0.148 | 0.240 | 0.261 |
| | 60 | Regression | 0.001 | 0.287 | 0.287 | 0.009 | 0.130 | 0.130 | -0.363 | 0.214 | 0.346 |
| | | Matching | 0.073 | 0.489 | 0.494 | 0.050 | 0.146 | 0.149 | -0.321 | 0.278 | 0.380 |
| | | C. matching | 0.012 | 0.548 | 0.548 | 0.012 | 0.153 | 0.153 | -0.386 | 0.315 | 0.464 |
| | | IPW | 0.111 | 1.172 | 1.184 | 0.088 | 0.457 | 0.465 | 0.135 | 0.458 | 0.476 |
| | | IIPW | 0.013 | 0.550 | 0.550 | 0.031 | 0.283 | 0.284 | -0.246 | 0.413 | 0.474 |
| | 80 | Regression | 0.027 | 0.589 | 0.589 | 0.015 | 0.180 | 0.180 | -0.593 | 0.490 | 0.842 |
| | | Matching | 0.142 | 1.076 | 1.096 | 0.075 | 0.189 | 0.194 | -0.507 | 0.483 | 0.740 |
| | | C. matching | 0.018 | 1.225 | 1.225 | 0.024 | 0.198 | 0.199 | -0.625 | 0.612 | 1.002 |
| | | IPW | 0.214 | 1.848 | 1.894 | 0.091 | 0.459 | 0.467 | 0.170 | 0.435 | 0.464 |
| | | IIPW | 0.005 | 0.997 | 0.997 | 0.029 | 0.290 | 0.291 | -0.503 | 0.483 | 0.735 |

**Table S4.2:** Performance of treatment effect estimators for various degrees $p$ of missing (MCAR) treatment indicator values and residual variances $\sigma^2$. *Abbreviations:* CCA, complete case analysis; $p$, missingness probability (%); PS method, propensity score method; $\bar{\hat{\beta}} - \beta$, estimated bias; Emp. var., empirical variance; MSE, empirical mean squared error; C. matching, calliper matching; IPW, inverse probability weighting; IIPW, iterative inverse probability weighting.

| | | | CCA | | | Within | | | Across | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma^2$ | $p$ | PS method | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE |
| 1 | 20 | Regression | 0.008 | 0.015 | 0.015 | -0.003 | 0.014 | 0.014 | -0.005 | 0.014 | 0.014 |
| | | Matching | 0.049 | 0.032 | 0.035 | 0.042 | 0.020 | 0.022 | 0.038 | 0.027 | 0.028 |
| | | C. matching | 0.003 | 0.032 | 0.032 | -0.002 | 0.018 | 0.018 | -0.006 | 0.026 | 0.026 |
| | | IPW | 0.041 | 0.347 | 0.348 | 0.017 | 0.253 | 0.254 | 0.018 | 0.247 | 0.247 |
| | | IIPW | 0.005 | 0.029 | 0.029 | -0.005 | 0.033 | 0.033 | -0.026 | 0.027 | 0.027 |
| | 40 | Regression | 0.004 | 0.019 | 0.019 | -0.026 | 0.015 | 0.016 | -0.031 | 0.015 | 0.016 |
| | | Matching | 0.048 | 0.044 | 0.046 | 0.017 | 0.019 | 0.019 | 0.014 | 0.024 | 0.024 |
| | | C. matching | -0.003 | 0.044 | 0.044 | -0.024 | 0.019 | 0.020 | -0.027 | 0.024 | 0.025 |
| | | IPW | 0.104 | 0.423 | 0.433 | 0.029 | 0.211 | 0.212 | 0.029 | 0.203 | 0.204 |
| | | IIPW | 0.001 | 0.054 | 0.054 | -0.022 | 0.039 | 0.039 | -0.059 | 0.028 | 0.032 |
| | 60 | Regression | -0.005 | 0.033 | 0.033 | -0.071 | 0.021 | 0.027 | -0.081 | 0.022 | 0.028 |
| | | Matching | 0.064 | 0.071 | 0.075 | -0.024 | 0.026 | 0.027 | -0.030 | 0.032 | 0.033 |
| | | C. matching | -0.005 | 0.070 | 0.070 | -0.067 | 0.025 | 0.030 | -0.072 | 0.031 | 0.036 |
| | | IPW | 0.147 | 0.545 | 0.567 | -0.005 | 0.165 | 0.165 | -0.006 | 0.154 | 0.154 |
| | | IIPW | 0.005 | 0.112 | 0.112 | -0.066 | 0.041 | 0.045 | -0.132 | 0.040 | 0.057 |
| | 80 | Regression | 0.015 | 0.065 | 0.065 | -0.163 | 0.033 | 0.060 | -0.186 | 0.034 | 0.069 |
| | | Matching | 0.096 | 0.160 | 0.169 | -0.108 | 0.040 | 0.051 | -0.111 | 0.042 | 0.054 |
| | | C. matching | 0.003 | 0.187 | 0.187 | -0.157 | 0.039 | 0.063 | -0.160 | 0.042 | 0.068 |
| | | IPW | 0.278 | 0.927 | 1.004 | -0.111 | 0.128 | 0.141 | -0.104 | 0.113 | 0.124 |
| | | IIPW | 0.020 | 0.170 | 0.171 | -0.160 | 0.045 | 0.070 | -0.257 | 0.076 | 0.142 |

Table S4.2 continued.

| | | | CCA | | | Within | | | Across | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma^2$ | $p$ | PS method | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE |
| 9 | 20 | Regression | 0.011 | 0.131 | 0.131 | 0.002 | 0.131 | 0.131 | -0.001 | 0.131 | 0.131 |
| | | Matching | 0.053 | 0.222 | 0.225 | 0.045 | 0.155 | 0.157 | 0.043 | 0.204 | 0.206 |
| | | C. matching | 0.004 | 0.228 | 0.228 | 0.001 | 0.156 | 0.156 | -0.002 | 0.209 | 0.209 |
| | | IPW | 0.099 | 0.528 | 0.538 | 0.074 | 0.409 | 0.415 | 0.077 | 0.402 | 0.408 |
| | | IIPW | 0.023 | 0.257 | 0.258 | 0.021 | 0.223 | 0.223 | 0.003 | 0.215 | 0.215 |
| | 40 | Regression | -0.021 | 0.183 | 0.184 | -0.046 | 0.185 | 0.187 | -0.052 | 0.185 | 0.188 |
| | | Matching | 0.026 | 0.327 | 0.328 | 0.001 | 0.217 | 0.217 | 0.001 | 0.245 | 0.245 |
| | | C. matching | -0.030 | 0.341 | 0.342 | -0.039 | 0.215 | 0.216 | -0.037 | 0.250 | 0.251 |
| | | IPW | 0.067 | 0.815 | 0.820 | 0.020 | 0.457 | 0.458 | 0.030 | 0.433 | 0.434 |
| | | IIPW | 0.012 | 0.541 | 0.541 | -0.030 | 0.273 | 0.273 | -0.068 | 0.271 | 0.276 |
| | 60 | Regression | 0.022 | 0.297 | 0.297 | -0.047 | 0.290 | 0.292 | -0.060 | 0.290 | 0.294 |
| | | Matching | 0.072 | 0.472 | 0.478 | -0.000 | 0.309 | 0.309 | -0.002 | 0.324 | 0.324 |
| | | C. matching | 0.004 | 0.532 | 0.532 | -0.046 | 0.314 | 0.316 | -0.047 | 0.331 | 0.333 |
| | | IPW | 0.146 | 1.021 | 1.042 | 0.016 | 0.470 | 0.471 | 0.037 | 0.449 | 0.450 |
| | | IIPW | 0.015 | 0.485 | 0.485 | -0.044 | 0.356 | 0.358 | -0.093 | 0.352 | 0.360 |
| | 80 | Regression | -0.017 | 0.549 | 0.549 | -0.161 | 0.501 | 0.527 | -0.188 | 0.504 | 0.540 |
| | | Matching | 0.125 | 1.011 | 1.026 | -0.100 | 0.527 | 0.536 | -0.106 | 0.538 | 0.549 |
| | | C. matching | 0.052 | 1.189 | 1.191 | -0.160 | 0.535 | 0.561 | -0.165 | 0.553 | 0.581 |
| | | IPW | 0.232 | 1.954 | 2.008 | -0.068 | 0.683 | 0.688 | -0.029 | 0.645 | 0.646 |
| | | IIPW | -0.023 | 1.029 | 1.029 | -0.130 | 0.600 | 0.617 | -0.209 | 0.584 | 0.627 |

**Table S4.3:** Performance of treatment effect estimators for various degrees $p$ of missing (MCAR) outcomes and residual variances $\sigma^2$. *Abbreviations:* CCA, complete case analysis; $p$, missingness probability (%); PS method, propensity score method; $\bar{\hat{\beta}} - \beta$, estimated bias; Emp. var., empirical variance; MSE, empirical mean squared error; C. matching, calliper matching; IPW, inverse probability weighting; IIPW, iterative inverse probability weighting.

| | | | CCA | | | Within | | | Across | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma^2$ | $p$ | PS method | $\bar{\hat{\beta}}-\beta$ | Emp. var. | MSE | $\bar{\hat{\beta}}-\beta$ | Emp. var. | MSE | $\bar{\hat{\beta}}-\beta$ | Emp. var. | MSE |
| 1 | 20 | Regression | 0.004 | 0.014 | 0.014 | 0.005 | 0.015 | 0.015 | 0.005 | 0.015 | 0.015 |
| | | Matching | 0.053 | 0.033 | 0.036 | 0.044 | 0.024 | 0.026 | 0.044 | 0.024 | 0.026 |
| | | C. matching | 0.010 | 0.033 | 0.033 | 0.006 | 0.024 | 0.024 | 0.006 | 0.024 | 0.024 |
| | | IPW | 0.084 | 0.324 | 0.331 | 0.059 | 0.261 | 0.265 | 0.059 | 0.261 | 0.265 |
| | | IIPW | -0.011 | 0.244 | 0.244 | 0.007 | 0.064 | 0.064 | 0.007 | 0.064 | 0.064 |
| | 40 | Regression | 0.004 | 0.020 | 0.020 | 0.003 | 0.022 | 0.022 | 0.003 | 0.022 | 0.022 |
| | | Matching | 0.054 | 0.049 | 0.052 | 0.039 | 0.034 | 0.035 | 0.039 | 0.034 | 0.035 |
| | | C. matching | 0.002 | 0.046 | 0.046 | 0.001 | 0.032 | 0.032 | 0.001 | 0.032 | 0.032 |
| | | IPW | 0.087 | 0.421 | 0.428 | 0.049 | 0.290 | 0.292 | 0.049 | 0.290 | 0.292 |
| | | IIPW | -0.002 | 0.074 | 0.074 | 0.011 | 0.101 | 0.101 | 0.011 | 0.101 | 0.101 |
| | 60 | Regression | 0.004 | 0.031 | 0.031 | 0.003 | 0.035 | 0.035 | 0.003 | 0.035 | 0.035 |
| | | Matching | 0.061 | 0.068 | 0.071 | 0.047 | 0.043 | 0.045 | 0.047 | 0.043 | 0.045 |
| | | C. matching | 0.000 | 0.073 | 0.073 | 0.010 | 0.043 | 0.043 | 0.010 | 0.043 | 0.043 |
| | | IPW | 0.171 | 0.524 | 0.553 | 0.056 | 0.291 | 0.294 | 0.056 | 0.291 | 0.294 |
| | | IIPW | 0.011 | 0.060 | 0.061 | -0.004 | 0.189 | 0.189 | -0.004 | 0.189 | 0.189 |
| | 80 | Regression | 0.003 | 0.073 | 0.073 | 0.000 | 0.082 | 0.082 | 0.000 | 0.082 | 0.082 |
| | | Matching | 0.116 | 0.166 | 0.179 | 0.040 | 0.092 | 0.094 | 0.040 | 0.092 | 0.094 |
| | | C. matching | 0.020 | 0.193 | 0.193 | -0.001 | 0.090 | 0.090 | -0.001 | 0.090 | 0.090 |
| | | IPW | 0.234 | 1.029 | 1.084 | 0.045 | 0.301 | 0.303 | 0.045 | 0.301 | 0.303 |
| | | IIPW | 0.015 | 0.300 | 0.301 | 0.002 | 0.087 | 0.087 | 0.002 | 0.087 | 0.087 |

Table S4.3 continued.

| | | | CCA | | | Within | | | Across | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma^2$ | $p$ | PS method | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE |
| 9 | 20 | Regression | -0.004 | 0.135 | 0.135 | -0.003 | 0.139 | 0.139 | -0.003 | 0.139 | 0.139 |
| | | Matching | 0.051 | 0.245 | 0.248 | 0.023 | 0.198 | 0.199 | 0.023 | 0.198 | 0.199 |
| | | C. matching | 0.007 | 0.254 | 0.254 | -0.016 | 0.199 | 0.200 | -0.016 | 0.199 | 0.200 |
| | | IPW | 0.065 | 0.499 | 0.503 | 0.035 | 0.450 | 0.451 | 0.035 | 0.450 | 0.451 |
| | | IIPW | -0.015 | 0.311 | 0.311 | -0.001 | 0.434 | 0.434 | -0.001 | 0.434 | 0.434 |
| | 40 | Regression | -0.002 | 0.181 | 0.181 | -0.004 | 0.199 | 0.199 | -0.004 | 0.199 | 0.199 |
| | | Matching | 0.054 | 0.333 | 0.336 | 0.031 | 0.255 | 0.256 | 0.031 | 0.255 | 0.256 |
| | | C. matching | -0.002 | 0.358 | 0.358 | -0.011 | 0.265 | 0.265 | -0.011 | 0.265 | 0.265 |
| | | IPW | 0.149 | 0.665 | 0.687 | 0.079 | 0.456 | 0.463 | 0.079 | 0.456 | 0.463 |
| | | IIPW | 0.003 | 0.380 | 0.380 | 0.006 | 0.283 | 0.283 | 0.006 | 0.283 | 0.283 |
| | 60 | Regression | 0.007 | 0.293 | 0.293 | 0.004 | 0.326 | 0.326 | 0.004 | 0.326 | 0.326 |
| | | Matching | 0.070 | 0.476 | 0.481 | 0.042 | 0.373 | 0.375 | 0.042 | 0.373 | 0.375 |
| | | C. matching | -0.007 | 0.548 | 0.549 | 0.004 | 0.381 | 0.381 | 0.004 | 0.381 | 0.381 |
| | | IPW | 0.170 | 1.119 | 1.148 | 0.077 | 0.675 | 0.681 | 0.077 | 0.675 | 0.681 |
| | | IIPW | 0.034 | 0.637 | 0.638 | 0.019 | 0.393 | 0.393 | 0.019 | 0.393 | 0.393 |
| | 80 | Regression | 0.007 | 0.597 | 0.597 | 0.024 | 0.698 | 0.699 | 0.024 | 0.698 | 0.699 |
| | | Matching | 0.139 | 1.055 | 1.074 | 0.053 | 0.723 | 0.726 | 0.053 | 0.723 | 0.726 |
| | | C. matching | 0.034 | 1.244 | 1.245 | 0.018 | 0.727 | 0.728 | 0.018 | 0.727 | 0.728 |
| | | IPW | 0.320 | 1.938 | 2.041 | 0.060 | 1.000 | 1.004 | 0.060 | 1.000 | 1.004 |
| | | IIPW | 0.036 | 1.115 | 1.116 | 0.004 | 0.824 | 0.824 | 0.004 | 0.824 | 0.824 |

**Table S4.4**: Performance of treatment effect estimators under various (MAR) missingness mechanisms and residual variances $\sigma^2$. *Abbreviations:* CCA, complete case analysis; MDM, missing data mechanism; PS method, propensity score method; $\bar{\hat{\beta}} - \beta$, estimated bias; Emp. var., empirical variance; MSE, empirical mean squared error; CP, empricial coverage probability; VR, ratio of mean estimated variance to empirical variance; C. matching, calliper matching; IPW, inverse probability weighting; IIPW, iterative inverse probability weighting. Under mechanism MAR1, the missingness of $X_2$ depends on $X_1$ and $T$ only. Under MAR2, the missingness depends on $Y$ only. Both MAR1 and MAR2 result in ∼40% incomplete records.

| | | | CCA | | | Within | | | Across | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma^2$ | MDM | PS method | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE |
| 1 | MAR1 | Regression | -0.003 | 0.016 | 0.016 | 0.002 | 0.014 | 0.014 | -0.059 | 0.016 | 0.020 |
| | | Matching | 1.054 | 0.075 | 1.187 | 0.042 | 0.018 | 0.020 | -0.084 | 0.029 | 0.036 |
| | | C. matching | -0.010 | 0.034 | 0.034 | 0.005 | 0.016 | 0.016 | -0.115 | 0.028 | 0.041 |
| | | IPW | -0.546 | 1.373 | 1.671 | 0.061 | 0.271 | 0.275 | 0.068 | 0.269 | 0.274 |
| | | IIPW | -0.003 | 1.128 | 1.128 | -0.002 | 0.036 | 0.036 | -0.021 | 0.024 | 0.025 |
| | MAR2 | Regression | -0.186 | 0.055 | 0.090 | -0.011 | 0.036 | 0.036 | -0.216 | 0.065 | 0.111 |
| | | Matching | -0.172 | 0.205 | 0.235 | 0.039 | 0.036 | 0.038 | -0.236 | 0.088 | 0.144 |
| | | C. matching | -0.181 | 0.210 | 0.243 | -0.004 | 0.039 | 0.039 | -0.303 | 0.105 | 0.197 |
| | | IPW | -0.005 | 0.187 | 0.187 | 0.038 | 0.283 | 0.284 | -0.003 | 0.307 | 0.307 |
| | | IIPW | -0.153 | 0.121 | 0.144 | -0.001 | 0.045 | 0.045 | -0.145 | 0.057 | 0.078 |

Table S4.4 continued.

| | | | CCA | | | Within | | | Across | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma^2$ | MDM | PS method | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE | $\bar{\hat{\beta}} - \beta$ | Emp. var. | MSE |
| 9 | MAR1 | Regression | -0.007 | 0.145 | 0.145 | 0.001 | 0.112 | 0.112 | -0.105 | 0.117 | 0.128 |
| | | Matching | 1.044 | 0.257 | 1.347 | 0.046 | 0.138 | 0.140 | -0.178 | 0.203 | 0.235 |
| | | C. matching | -0.001 | 0.297 | 0.297 | 0.011 | 0.141 | 0.141 | -0.201 | 0.214 | 0.254 |
| | | IPW | -0.532 | 2.143 | 2.425 | 0.081 | 0.409 | 0.415 | 0.093 | 0.407 | 0.416 |
| | | IIPW | -0.003 | 2.358 | 2.358 | 0.001 | 0.210 | 0.210 | -0.026 | 0.208 | 0.209 |
| | MAR2 | Regression | -0.812 | 0.225 | 0.885 | -0.014 | 0.147 | 0.147 | -0.394 | 0.234 | 0.389 |
| | | Matching | -0.745 | 0.506 | 1.061 | 0.038 | 0.166 | 0.167 | -0.362 | 0.305 | 0.436 |
| | | C. matching | -0.746 | 0.516 | 1.072 | -0.016 | 0.170 | 0.170 | -0.487 | 0.388 | 0.625 |
| | | IPW | -0.540 | 0.542 | 0.833 | 0.016 | 0.472 | 0.472 | -0.134 | 0.544 | 0.562 |
| | | IIPW | -0.654 | 0.570 | 0.998 | -0.024 | 0.234 | 0.234 | -0.440 | 0.318 | 0.512 |

Table S4.4 continued.

| | | | | Within | | Across | |
|---|---|---|---|---|---|---|---|
| $\sigma^2$ | MDM | PS method | Variance estimation/95%CI method | CP | VR | CP | VR |
| 1 | MAR1 | Regression | Bootstrapping | 0.951 | 1.046 | 0.929 | 1.041 |
| | | | Rubin's rules | 0.975 | 1.378 | 0.899 | 0.783 |
| | | Matching | Bootstrapping | 0.969 | 1.234 | 0.959 | 1.435 |
| | | C. matching | Bootstrapping | 0.967 | 1.287 | 0.944 | 1.607 |
| | | IPW | Bootstrapping | 0.845 | 0.976 | 0.839 | 0.975 |
| | | IIPW | Bootstrapping | 0.941 | 2.768 | 0.935 | 5.319 |
| | MAR2 | Regression | Bootstrapping | 0.956 | 1.197 | 0.878 | 1.249 |
| | | | Rubin's rules | 0.971 | 1.330 | 0.484 | 0.224 |
| | | Matching | Bootstrapping | 0.964 | 1.252 | 0.894 | 1.372 |
| | | C. matching | Bootstrapping | 0.959 | 1.297 | 0.839 | 1.409 |
| | | IPW | Bootstrapping | 0.869 | 0.980 | 0.888 | 0.992 |
| | | IIPW | Bootstrapping | 0.959 | 1.873 | 0.895 | 2.242 |
| 9 | MAR1 | Regression | Bootstrapping | 0.957 | 1.035 | 0.940 | 1.039 |
| | | | Rubin's rules | 0.957 | 1.076 | 0.923 | 0.914 |
| | | Matching | Bootstrapping | 0.968 | 1.192 | 0.977 | 1.580 |
| | | C. matching | Bootstrapping | 0.969 | 1.201 | 0.976 | 1.704 |
| | | IPW | Bootstrapping | 0.896 | 1.034 | 0.893 | 1.034 |
| | | IIPW | Bootstrapping | 0.947 | 1.126 | 0.945 | 1.503 |
| | MAR2 | Regression | Bootstrapping | 0.948 | 1.057 | 0.893 | 1.152 |
| | | | Rubin's rules | 0.953 | 1.081 | 0.671 | 0.453 |
| | | Matching | Bootstrapping | 0.936 | 1.150 | 0.872 | 1.525 |
| | | C. matching | Bootstrapping | 0.910 | 1.181 | 0.800 | 1.568 |
| | | IPW | Bootstrapping | 0.927 | 1.035 | 0.941 | 1.070 |
| | | IIPW | Bootstrapping | 0.960 | 1.959 | 0.880 | 1.596 |

## References

Mitra, R. and J. P. Reiter (2016): "A comparison of two methods of estimating propensity scores after multiple imputation," *Statistical methods in medical research*, 25, 188–204.

Wan, F. and N. Mitra (2016): "An evaluation of bias in propensity score-adjusted non-linear regression models," *Statistical methods in medical research*, 0, 1–17.

# 5

# Propensity score estimation using classification and regression trees in the presence of missing covariate data

Bas B. L. Penning de Vries
Maarten van Smeden
Rolf H. H. Groenwold

*Epidemiologic Methods* 2018; 7(1)

Propensity score estimation, CART and missing covariates

## Abstract

Data mining and machine learning techniques such as classification and regression trees (CART) represent a promising alternative to conventional logistic regression for propensity score estimation. Whereas incomplete data preclude the fitting of a logistic regression on all subjects, CART is appealing in part because some implementations allow for incomplete records to be incorporated in the tree fitting and provide propensity score estimates for all subjects. Based on theoretical considerations, we argue that the automatic handling of missing data by CART may however not be appropriate. Using a series of simulation experiments, we examined the performance of different approaches to handling missing covariate data; (i) applying the CART algorithm directly to the (partially) incomplete data, (ii) complete case analysis, and (iii) multiple imputation. Performance was assessed in terms of bias in estimating exposure-outcome effects among the exposed, standard error, mean squared error and coverage. Applying the CART algorithm directly to incomplete data resulted in bias, even in scenarios where data were missing completely at random. Overall, multiple imputation followed by CART resulted in the best performance. Our study showed that automatic handling of missing data in CART can cause serious bias and does not outperform multiple imputation as a means to account for missing data.

## 5.1 Introduction

Propensity score analysis has gained increasing popularity as means to adjust for measured confounding (Rosenbaum and Rubin, 1983; Stürmer et al., 2006). Inference typically proceeds by stratification on the propensity score, propensity score adjustment in a regression model, inverse probability weighting (IPW) or matching based on propensity scores given measured covariates (Rosenbaum and Rubin, 1983; Austin, 2011a). It is standard practice to obtain estimates of the propensity score by a parametric (logistic) regression of the exposure on measured covariates. However, parametric models rely on assumptions about the distribution of variables in relation to one another, including the functional form and the presence or absence of interactions. If any of these are violated, covariate balance may not be attained, potentially leading to bias in making causal inferences about the exposure-outcome relation of interest (Drake, 1993).

It has been suggested that machine learning and data mining methods, such as classification and regression tree analysis (CART), be used to estimate the relationship between the exposure and measured covariates. These methods avoid making the assumptions regarding functional form and interaction as in a standard logistic regression. The utility of data mining methods to estimate propensity scores in complete data settings has been studied previously (Setoguchi et al., 2008; Lee et al., 2010; Westreich et al., 2010; Wyss et al., 2014). However, in practice, researchers are often faced with missing values on the measured variables. Whereas incomplete data preclude logistic regression on all subjects, some CART algorithms allow for incomplete records to be incorporated in the tree fitting and provide propensity score estimates for all subjects. The ability of CART to accommodate missing values has been described as advantageous (Lee et al., 2010; McCaffrey et al., 2004; Moisen, 2008; Rai et al., 2017). However, the precise impact of missing data on the performance of CART-based propensity score estimators has received little attention. The objective of this study was therefore to examine the performance of various CART-based propensity score estimation procedures in the presence of missing data. Throughout, particular emphasis is placed on the causal odds ratio for the marginal effect among the exposed (or Average Effect among the 'Treated', ATT) as the effect measure of interest.

The remainder of this article is structured as follows. In Section 5.2, we briefly review pertinent theory. Based on analytical work, we identify caveats in the handling of missing data by CART. Section 5.3 describes a series of Monte Carlo simulations that were used to evaluate the performance of various approaches to handling missing data, including (i) subjecting incomplete data directly to the

CART algorithm, (ii) complete case analysis, and (iii) multiple imputation. In Section 5.4, we apply and compare the approaches in a case study on the effect of influenza vaccination and mortality. We conclude with a summary and discussion of our findings in the context of the existing literature.

## 5.2 Theory

### *5.2.1 Propensity score analysis of complete data*

*Counterfactual outcomes and estimating causal effects*

We adopt a perspective of potential or counterfactual outcomes, formal accounts of which are given for example by Neyman et al. (1935), Rubin (1974), Holland (1986), Holland (1988) and Pearl (2009).

Consider a sequence $S = (X_1, X_2, ..., X_n)$ of variables and let $\mathcal{F} = (f_{X_1}, f_{X_2}, ..., f_{X_n})$ be a collection of functions $f_{X_j}$ that deterministically map a realisation of the predecessors $(X_i : i < j)$ of $X_j$ and of exogenous variable $\varepsilon_{X_j}$ into a realisation of $X_j$. We may write the random variables $X_1, X_2, ..., X_n$ as follows:

$$\left.\begin{aligned} X_1 &= f_{X_1}(\varepsilon_{X_1}), \\ X_2 &= f_{X_2}(X_1, \varepsilon_{X_2}), \\ &\vdots \\ X_n &= f_{X_n}(X_1, X_2, ..., X_{n-1}, \varepsilon_{X_n}). \end{aligned}\right\} \tag{5.1}$$

Now, for any intervention setting $X_t$ to $x_t$ for all $t$ in a subset $T$ of $\{1, 2, ..., n\}$, the counterfactual versions of $X_1, X_2, ..., X_n$ are obtained by applying (5.1) with $X_t = f_X(X_1, ..., X_{t-1}, \varepsilon_{X_t})$ replaced with $X_t = x_t$ if $t \in T$.

Specifically, let $S = (W, A, Y, R)$ so that $W = f_W(\varepsilon_W)$, $A = f_A(W, \varepsilon_A)$, $Y = f_Y(W, A, \varepsilon_Y)$, and $R = f_R(W, A, Y, \varepsilon_R)$. $W$ may be thought of as a (random vector of) baseline or pre-exposure variable(s), $A$ denotes the binary exposure of interest, $Y$ the outcome, and $R$ a missing indicator vector of $W$. A subject's counterfactual outcomes $Y_0$ and $Y_1$, obtained if exposure $A$ were set possibly contrary to fact to 0 and 1, respectively, are defined such that $Y_0 = f_Y(W, 0, \varepsilon_Y)$ and $Y_1 = f_Y(W, 1, \varepsilon_Y)$.

Causal effects are readily defined in terms of counterfactual outcomes. In this article, the focus is on the causal odds ratio (OR) for the marginal effect of exposure $A$ on binary outcome $Y$ among the exposed ($A = 1$), that is,

$$\mathrm{OR} = \frac{\mathbb{E}[Y_1|A=1]/(1-\mathbb{E}[Y_1|A=1])}{\mathbb{E}[Y_0|A=1]/(1-\mathbb{E}[Y_0|A=1])}.$$

Under consistency as defined by Cole and Frangakis (2009), $Y_1$ is equal to the observed outcome $Y$ for subjects in the exposure group. $Y_0$, on the other hand, is not observed for exposed subjects. We may, however, validly estimate the causal OR under a set of conditions, which includes no interference between subjects (or Stable Unit Treatment Value Assumption, Tchetgen and VanderWeele, 2012), consistency, positivity, and conditional exchangeability (Lesko et al., 2017). To simplify arguments and notation, we shall assume that all of these conditions hold, with the exception of conditional exchangeability, unless otherwise indicated. If there exists a (set of) variable(s) $Z$ such that the potential outcomes are conditionally independent of exposure status given $Z$, we may write

$$\begin{aligned} P(Y_0|A=1) &= \mathbb{E}[P(Y_0|A=1,Z)|A=1] \\ &= \mathbb{E}[P(Y_0|A=0,Z)|A=1] \\ &= \mathbb{E}[P(Y|A=0,Z)|A=1], \end{aligned}$$

so that the causal OR may be expressed entirely in terms of observables

$$\text{OR} = \frac{\mathbb{E}[Y|A=1]/(1-\mathbb{E}[Y|A=1])}{\mathbb{E}\{\mathbb{E}[Y|A=0,Z]|A=1\}/(1-\mathbb{E}\{\mathbb{E}[Y|A=0,Z]|A=1\})}.$$

$W$ satisfies the definition of $Z$ whenever $\varepsilon_Y \perp\!\!\!\perp \varepsilon_A|W$. In practice, validly estimating $\mathbb{E}[Y|A=0,Z]$ may be difficult when $Z$ is multidimensional and $Y$ is rare (Albert and Anderson, 1984). In this case, it may be desirable to summarise $Z$ in a single balancing score (Rosenbaum and Rubin, 1983).

### *The propensity score*

The propensity score $e(W)$, defined as the conditional probability of exposure given covariates $W$, satisfies a number of balancing properties. First, covariate(s) $W$ and exposure $A$ are conditionally independent given the propensity score, and conditional exchangeability given covariate(s) $W$ implies conditional exchangeability given $e(W)$ (Rosenbaum and Rubin, 1983, Theorems 1 and 3). Thus, the causal OR becomes

$$\text{OR} = \frac{\mathbb{E}[Y|A=1]/(1-\mathbb{E}[Y|A=1])}{\mathbb{E}\{\mathbb{E}[Y|A=0,e(W)]|A=1\}/(1-\mathbb{E}\{\mathbb{E}[Y|A=0,e(W)]|A=1\})}.$$

This formulation has motivated the propensity score matching approach as discussed by Rosenbaum and Rubin (1983).

Balance may also be attained by inverse probability weighting (Appendix A). To simplify arguments and notation, we assume that $W$ and $Y$ take a discrete joint distribution; however, the results extend to continuous or mixed discrete/continuous distributions. To obtain an IPW estimator of the ATT, let

$$\varphi(w,a) = \frac{\varphi^*(w,a)}{\mathbb{E}[\varphi^*(W,A)|A=a]}, \qquad \varphi^*(w,a) = I(a=1) + I(a=0)\frac{e(w)}{1-e(w)},$$

for realisations $w$ of $W$ and $a$ of $A$, where $I$ denotes the indicator function taken the value 1 if the argument is true and 0 otherwise. Weighting by $\varphi$ yields independence between covariate(s) $W$ and $A$; that is, for all $w$,

$$\varphi(w,0)\Pr(W=w|A=0) = \varphi(w,1)\Pr(W=w|A=1).$$

Also, conditional exchangeability given $W$ implies exchangeability following weighting by $\varphi$; that is, if $(Y_0,Y_1) \perp\!\!\!\perp A|W=w$ for all $w$, then

$$\begin{aligned}&\sum_w \varphi(w,0)\Pr(Y_0=y_0,Y_1=y_1,W=w|A=0)\\ &\qquad\qquad = \sum_w \varphi(w,1)\Pr(Y_0=y_0,Y_1=y_1,W=w|A=1)\end{aligned}$$

for all $y_0, y_1$. Thus, the causal OR becomes

$$\begin{aligned}\mathrm{OR} = &\frac{\sum_w \varphi(w,1)\Pr(Y=1,W=w|A=1)}{\{1-\sum_w \varphi(w,1)\Pr(Y=1,W=w|A=1)\}}\\ &\bigg/ \frac{\sum_w \varphi(w,0)\Pr(Y=1,W=w|A=0)}{\{1-\sum_w \varphi(w,0)\Pr(Y=1,W=w|A=0)\}}.\end{aligned}$$

In words, this means that the causal odds ratio is equal to the crude odds ratio of the ATT in the (pseudo-)population that is obtained by weighting each observation by $\varphi$.

### *Ensemble CART methods in the absence of missing data*

We will now briefly describe how CART can be applied to estimate the propensity score. Detailed information can be found elsewhere (McCaffrey et al., 2004; Breiman, 1996; Ridgeway, 1999; Breiman, 2001; Elith et al., 2008; Hastie et al., 2009). CART is a type of supervised learning task that entails finding a set of rules, subject to constraints, that partition the data into regions based on the input data (covariates) such that within regions, target values (e.g., exposure levels) meet a desirable level of homogeneity. Typically, a tree is built in a recursive manner by splitting the dataset into increasingly homogeneous subsets and choosing the splitting rule at each step or node that best splits the data further, with ‘best’ referring to the greatest improvement in terms of some homogeneity metric, such as the Gini index (Therneau and Atkinson, 2017). Ensemble techniques by definition fit more than one tree to the data and combine them to form a single predictor of ‘the outcome’ (in the case of propensity score, the assigned exposure). The aim of ensemble techniques is to enhance performance and reduce issues of overfitting by a single tree (Elith et al., 2008; Moisen, 2008; Hastie et al., 2009). We focus here on two popular CART ensemble methods, namely boostrap aggregated (bagged) CART and boosted CART.

*Bootstrap aggregated CART.* Bagged CART involves drawing bootstrap samples form the original study sample (Breiman, 1996). A CART tree is formed in each bootstrap sample, yielding multiple predictors of the target variable. For each subject, the final prediction is formed by the average or majority vote across all predictors. In the context of propensity scores, the prediction of a single tree for any given subject may be defined as the proportion of exposed subjects among those individuals that are assigned to the same region by the given tree. The final propensity score is the average of the predictions across all bootstrap samples. Propensity score matching may then be thought of as matching exposed subjects to unexposed subjects from the same or ‘nearby’ region.

*Boosted CART.* Boosted CART is related to bagged CART in the sense that it is an ensemble method; multiple trees are fit and merged to form a single predictor. With boosted CART, trees are fit in a forward, stagewise procedure. In boosting, trees are fit iteratively to the data such those observations whose observed exposure levels are poorly predicted by the predictor of the previous iteration receive greater weight at the current iteration (Ridgeway, 1999; Elith et al., 2008). Some implementations construct trees using data splits aimed not at achieving homogeneity of the exposure values themselves, but at achieving homogeneity of prediction error of the estimator obtained in the previous step (McCaffrey et al., 2004; Elith et al., 2008). With each iteration, a new predictor is formed by making adjustments to the predictor obtained in the previous step. The final predictor is constructed with contributions from all trees.

### *5.2.2 Ignorable missing data and generalised propensity scores*

In this section, we briefly review the concept of ignorable missing data, and discuss a generalisation of the propensity score which allows for missing data as well as strategies to incorporate missing data directly in the CART fitting. For certain CART algorithms (in our case boosted CART), the inherent missing data strategy yields estimates of the generalised propensity score.

### *Ignorable missing data*

Suppose $W = (W_1, W_2, ..., W_p)$ and $R = (R_1, R_2, ..., R_p)$ are random vectors of size $p$ such that for $j = 1, 2, ..., p$, $R_j = 0$ if $W_j$ is missing and $R_j = 1$ if $W_j$ is observed. Following Rubin (1976), define the extended random vector $V = (V_1, V_2, ..., V_p)$ with range to include the special value $*$ to indicate a missing datum: $V_j = W_j$ if $R_j = 1$ and $V_j = *$ if $R_j = 0$. Let $v$ be a particular sample realisation of $V$, so that each $v_j$ is either a known quantity or $*$. These values imply a realisation for the random variable $R$, denoted $r$. For notational convenience, we write $W = (W^{\text{obs}}, W^{\text{mis}})$ and $V = (V^{\text{obs}}, V^{\text{mis}})$ to indicate that each may be partitioned into two vectors corresponding to all $j$ such that $r_j = 1$ for observed data and $r_j = 0$ for missing data. It is important to note that these partitions are defined with respect to $r$, the observed pattern of missing data. Given a realisation $r$ of $R$, and provided that $A, Y$ are observed, covariate data are said to be missing at random (MAR) if $\Pr(R = r | W^{\text{obs}}, W^{\text{mis}} = u, A, Y)$ and $\Pr(R = r | W^{\text{obs}}, W^{\text{mis}} = u', A, Y)$ are the same for all $u, u'$ and at each possible value of the parameter vector $\phi$ that fully characterises the missing data mechanism (Rubin, 1976). If in addition to MAR, the parameter $\phi$ is distinct, in the sense of Rubin (1976), from the vector $\theta$ that parameterises the distribution of the data that we would have based inference on had there been no missingness, then missing data is said to be ignorable and it is not necessary to consider the missing data or the missing data mechanism in making inferences about $\theta$ (Rubin, 1976, 1987; Schafer, 1997). Thus, if the missing data mechanism is ignorable, one may validly model the complete data to create imputations for the missing data (Rubin, 1987; Van Buuren, 2012).

### *The generalised propensity score*

The generalised propensity score $e^*(V)$ is defined as the conditional exposure probability given the extended covariate vector $V$ (D'Agostino Jr. and Rubin, 2000). That is,

$$\begin{aligned} e^*(V) &= \Pr(A = 1 | W^{\text{obs}}, R) \\ &= \sum_w \Pr(A = 1 | W, R) \Pr(W^{\text{mis}} = w | W^{\text{obs}}, R). \end{aligned}$$

Using the same argumentation to establish the balancing properties of the usual propensity score, it can be shown that the generalised propensity score has the same balancing properties with respect to $V$ as the usual propensity score has with respect to $W$. Thus, the observed covariate data and missingness information and exposure $A$ are conditionally independent given the generalised propensity score, and conditional exchangeability given the extended covariate(s) $V$ implies conditional exchangeability given the generalised propensity score $e^*(V)$.

To obtain an IPW estimator of the ATT, let

$$\gamma(v,a) = \frac{\gamma^*(v,a)}{\mathbb{E}[\gamma^*(V,A)|A=a]}, \qquad \gamma^*(v,a) = I(a=1) + I(a=0)\frac{e^*(v)}{1-e(v)},$$

for realisations $v$ of $V$ and $a$ of $A$, Then, weighting by $\gamma$ renders $V$ independent of $A$; that is, for all $v$,

$$\gamma(v,0)\Pr(V=v|A=0) = \gamma(v,1)\Pr(V=v|A=1).$$

Also, conditional exchangeability given $V$ implies conditional exchangeability following weighting by $\gamma$; that is, if $(Y_0,Y_1) \perp\!\!\!\perp A|V$, then

$$\sum_v \gamma(v,0)\Pr(Y_0=y_0,Y_1=y_1,V=v|A=0) \\ = \sum_v \gamma(v,1)\Pr(Y_0=y_0,Y_1=y_1,V=v|A=1)$$

for all $y_0, y_1$.

Importantly, the propensity score $e(W)$ need not equal the generalised propensity score $e^*(V)$. That is, given the observed covariate data, the unobserved covariate data need not provide the same information about exposure allocation as does the missing data pattern. In addition, neither covariate balance given the propensity score ($W \perp\!\!\!\perp A|e(W)$) nor balance of the observed data and missingness information given the generalised propensity score ($V \perp\!\!\!\perp A|e^*(V)$) generally implies covariate balance given the generalised propensity score ($W \perp\!\!\!\perp A|e^*(V)$).

More crucially perhaps, conditional exchangeability given the generalised propensity score is not guaranteed even if conditional exchangeability given the usual propensity score holds or the generalised propensity score balances both observed and unobserved covariate data (i.e., neither $(Y_0,Y_1) \perp\!\!\!\perp A|e(W)$ nor $W \perp\!\!\!\perp A|e^*(V)$ nor both imply that $(Y_0,Y_1) \perp\!\!\!\perp A|e^*(V)$; see Appendix B for an example).

This suggests that it is not generally desirable to distribute across exposure groups both the observed data and the missingness information by adjusting for the generalised propensity score. However, there are situations conceivable in which it is appropriate to base inference on the generalised rather than the usual propensity score. Until now, we have assumed an ordering of the variables in which the outcome $Y$ precedes $R$, the missingness pattern of $W$. Consequently, $Y$ was defined as a function $f_Y$ of $W$, $A$ and exogenous variable $\varepsilon_Y$ and not of $R$. Consider now a setting where $S = (W, R, A, Y)$ so that $R$ forms a predecessor of $A$ and $Y$ (and, therefore, a potential common cause of $A$ and $Y$). Then, if exchangeability can be attained by conditioning on $e^*(V)$, conditional exchangeability given $e(W)$ need not hold (see Appendix C for an example).

Thus, the choice between adjustment for the generalised versus the usual propensity score should ideally rest on the relative extent to which conditional exchangeability holds given the generalised versus the usual propensity score. In practice, it is not possible to estimate directly the true propensity score when covariate data are missing (Rosenbaum and Rubin, 1983; D'Agostino Jr. and Rubin, 2000; Cham and West, 2016). However, under ignorability of missing data, one may 'recover' the unobserved data, e.g., via multiple imputation (Rubin, 1987; Van Buuren, 2012), prior to estimating propensity scores. Henceforth, we assume that exchangeability can be attained by conditioning on the complete covariate data or, therefore, the usual propensity score, if data were not missing. We also assume that missing data is ignorable.

### *Applying CART to incomplete data*

*Bootstrap aggregated CART.* In this study, we used bagged CART as implemented in the R package `ipred` (Peters and Hothorn, 2017, version 0.9-6). This implementation allows for missing data by first evaluating homogeneity at a given node among only those observations whose candidate splitting variable is observed. Once the splitting variable and split point have been decided, the algorithm uses a surrogate splits approach to classify records whose splitting variable is missing based on the other variables included in the tree fitting (Therneau and Atkinson, 2017).

The bagged CART algorithm replaces missing confounder values without regard of the outcome or exposure status. As a result, any two subjects whose covariate data are identical, except possibly for the missing covariate, would be allocated to the same covariate region by any given tree. However, subjects within a given region need not be exchangeable. In fact, systematic differences in the outcome of the causal model ($Y$) between exposed and unexposed subjects may be in part attributable to the missing covariate (confounder). As such, even under completely at random missingness (MCAR), we would expect propensity score matching or IPW based on bagged CART to yield bias in the direction of confounding by the missing covariate.

*Boosted CART.* An implementation of boosted CART to estimate propensity scores is available in the R package `twang` (Ridgeway et al., 2017, version 1.5). This implementation allows for incomplete records to be incorporated in the tree fitting by regarding missingness as a special covariate level and assigning to a given (non-terminal) node three child nodes; one to which any individual is allocated whose splitting variable is missing, one for observed values that exceed some threshold, and one for the remainder. That is, rather than modelling the relationship between exposure and covariates, an attempt is made to model the association between exposure on the one hand and observed covariate data and missingness information on the other hand, and, therefore, to construct scores that balance the missingness across the matched or weighted exposure groups. In other words, the algorithm represents an estimator of the generalised propensity score.

While boosted CART may be successful at distributing missingness rates across exposure groups, it makes no attempt at distributing the unobserved values. If the partially observed covariate represents a confounder, systematic differences across exposure groups may persist after propensity score matching or IPW based on the generalised propensity score. As such, under MCAR, we would expect boosted CART to yield a propensity score matching or IPW estimator that is biased in the direction of confounding by the partially observed covariate. When missingness is MAR dependent on the outcome, boosted CART tends to render exposure groups more comparable in terms of the outcome and, therefore, attenuate the apparent exposure-outcome effect.

*Bias when applying CART to incomplete data.* In summary, above, we argued that using either boosted CART or bagged CART to estimate propensity scores may yield a biased estimator of the causal ATT, when applying the CART algorithm directly to the (partially) incomplete data. In bagged CART, missing confounder values are replaced, yet this procedure may not be appropriate, since exposure and outcome status are ignored in this process. Boosted CART, on the other hand, balances observed covariate values as well as missing indicator values. Since the latter may depend on the outcome (under the assumption of ignorability), boosted CART potentially balances outcome values too, yielding a biased estimator of the causal effect.

## 5.3 Monte Carlo simulations

We now describe a simulation study in which we evaluated the performance of CART-based propensity score matching and IPW in the presence of missing confounder data.

### 5.3.1 Methods

*Simulation structure*

We performed a series of Monte-Carlo simulation experiments based on the simulation structure described in Setoguchi et al. (2008) with modifications so as to allow for missing data. For $n = 2000$ subjects, we independently generated 10 covariates $W_i$ (four confounders, three predictors of the exposure only, and three predictors of the outcome only), a binary exposure variable $A$, and a binary outcome $Y$ (Figure 5.1). Missing data were introduced into one or two covariates. A number of CART-based approaches were used to estimate propensity scores, before and after the introduction of missing data, and in turn the log odds ratio for the exposure-outcome effect among the treated. For comparison, we also estimated propensity scores in imputed datasets using a correctly specified propensity score model, and using a logistic model with main effects only. The process was repeated 5000 times for each of eight simulation scenarios that varied primarily by missing data mechanism. All simulations were conducted with R-3.2.2 on a Windows 7 (64-bit) platform (R Core Team, 2016).

*Data generation*

Data were generated by sequentially going through the following steps. First, the covariates were generated by sampling from a multivariate normal distribution

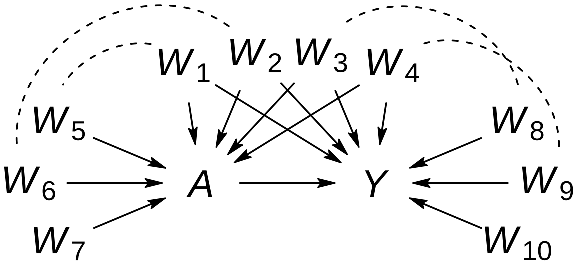


**Figure 5.1:** Complete data structure for simulation experiments. Dashed arcs without arrowheads connecting variables indicate non-zero entries for the corresponding variables in the covariance matrix of the joint distribution of all $W_i$, $i = 1, 2, ..., 10$.

with zero means and unit variances; correlations were set to zero except for the correlations between $W_1$ and $W_5$, $W_2$ and $W_6$, $W_3$ and $W_8$, and $W_4$ and $W_9$, which were set to 0.2, 0.9, 0.2, and 0.9, respectively. Second, covariates $W_1$, $W_3$, $W_5$, $W_6$, $W_8$, and $W_9$ were dichotomised, setting any value to 1 if greater than 0 and to 0 otherwise.

Following Setoguchi et al. (2008), the binary exposure variable $A$ was related to the covariate vector following the propensity score model $\Pr(A = 1|W) = \text{expit}\{0.8W_1 - 0.25W_2 + 0.6W_3 - 0.4W_4 - 0.8W_5 - 0.5W_6 + 0.7W_7 - 0.25W_2^2 - 0.4W_4^2 + 0.7W_7^2 + 0.4W_1W_3 - 0.175W_2W_4 + 0.3W_3W_5 - 0.28W_4W_6 - 0.4W_5W_7 + 0.4W_1W_6 - 0.175W_2W_3 + 0.3W_3W_4 - 0.2W_4W_5 - 0.4W_5W_6\}$. Realisations $a$ for $A$ were generated by drawing a pseudo-value from the uniform(0,1) distribution and setting $a$ to 1 if this number was less than the true propensity score and to 0 otherwise. Consequently, $A$ can be thought of as a exposure that is generated by a non-linear and non-additive propensity score model. This model assigns approximately half of subjects to the exposure group.

Outcomes were generated following the mechanism described by Setoguchi et al. (2008) with slight modifications to increase the outcome fraction (from approximately 2% to 20% or 40%). Specifically, the binary outcome, $Y$, was modelled as a Bernoulli random variable given $A$ and $W$: an independent random number, $\varepsilon_Y$, was drawn from the uniform distribution; $Y$ was set to 1 if this number was less than the inverse logit (expit) of a linear transformation $\eta(A, W) = -1 + 0.3W_1 - 0.36W_2 - 0.73W_3 - 0.2W_4 + 0.71W_8 - 0.19W_9 + 0.26W_{10} + \gamma A$ of $A$ and $W$ and to 0 otherwise. The true conditional log odds ratio for the exposure-outcome effect was set to 1 or $-1$ depending on the scenario. The outcome incidence was roughly 40% for scenarios with $\gamma = 1$; and 20% for scenarios with $\gamma = -1$. The counterfactual outcomes $Y_0$ and $Y_1$ for any subject with realisations $w$ of $W$ and $u$ of $\varepsilon_Y$ are found by computing $I(u < \text{expit}\{\eta(0, w)\})$ and $I(u < \text{expit}\{\eta(1, w)\})$, $I$ denoting the indicator function. With knowledge of the counterfactual outcomes, it can be inferred that with $\gamma = 1$, the marginal log odds ratio for the true exposure-outcome effect among the exposed (or treated; ATT) is approximately 0.906; with $\gamma = -1$, the marginal log odds ratio is approximately $-0.926$ (Hernán and Robins, 2017). Note that these are different from the conditional causal odds ratios as a result of the non-collapsibility property of the odds ratio.

We considered ignorable missing data mechanisms for introducing missing data.

*MCAR missingness.* For all subjects, irrespective of complete data, values of $W_3$ were set to missing with probability $p$, characterising the MCAR mechanism. The missingness probability of the other variables was set to zero.

*MAR missingness.* Let $M_3$ be a missing indicator variable that takes the value of one if and only if the value of $W_3$ is missing. Similarly, define $M_4$ to be the missing indicator variable pertaining to $W_4$. Given the full data, $W_3$ and $W_4$ were set to missing independently of one and other and with probability equal to $\Pr(M_3 = 1|W, A, Y) = p$ and $\Pr(M_4 = 1|W, A, Y) = \text{expit}\{\alpha_0 + \alpha_1 W_1 + \alpha_2 A + \alpha_3 Y\}$. The missingness probability of the other variables was set to zero.

### *Scenarios*

We evaluated the performance of various CART-based methods in eight scenarios (Table 5.1). The intercepts $\alpha_0$ in scenarios five through eight were chosen so as to yield roughly the same average proportion of missing data points per generated dataset of 24000 data points (2000 records on 10 covariates, one exposure and one outcome variable), namely 3%. The average proportion of missing data points and the fraction of incomplete records were largest in scenario 2 (5% and 60%, respectively; see Table 5.1). In all of the scenarios considered, data are 'missing at random' and it is assumed that there is conditional exchangeability given measured covariates (i.e., $(Y_0, Y_1) \perp\!\!\!\perp A|W$).

Note that in scenarios 3 trough 8, conditioning on $M$ may break the independence between $A$ and unobserved outcome predictor $\varepsilon_Y$ through what is known as collider stratification (cf. Pearl, 2009). One might therefore expect that that discarding incomplete records in these scenarios would result in bias. In

| Scenario | $\gamma$ | MCAR/MAR | $p$ | $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | PMP | PIR |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | MCAR | 0.3 | – | – | – | – | 0.03 | 0.30 |
| 2 | 1 | MCAR | 0.6 | – | – | – | – | 0.05 | 0.60 |
| 3 | 1 | MAR | 0.0 | −0.7 | 0.0 | 0.0 | 1.5 | 0.04 | 0.48 |
| 4 | −1 | MAR | 0.0 | −1.0 | 0.0 | 0.0 | 1.5 | 0.03 | 0.35 |
| 5 | 1 | MAR | 0.1 | −1.6 | 0.5 | 0.5 | 0.5 | 0.03 | 0.37 |
| 6 | 1 | MAR | 0.1 | −2.1 | 0.5 | 0.5 | 1.5 | 0.03 | 0.37 |
| 7 | 1 | MAR | 0.1 | −2.3 | 0.5 | 1.5 | 0.5 | 0.03 | 0.36 |
| 8 | 1 | MAR | 0.1 | −2.2 | 1.5 | 0.5 | 0.5 | 0.03 | 0.37 |

**Table 5.1:** Description of scenarios. $\gamma$ equals the conditional log odds ratio for the effect of $A$ on $Y$ given $W$. Given the full data, variables $W_3$ and $W_4$ were set to missing independently of one and other and with probabilities $p$ and $\text{expit}\{\alpha_0 + \alpha_1 W_1 + \alpha_2 A + \alpha_3 Y\}$, respectively. Abbreviations: MCAR, missing completely at random; MAR, missing at random; PMP, average proportion of missing data points; PIR, average proportion of incomplete records.

scenarios 3 and 4, however, covariate missingness $M$ is conditionally independent of exposure status and covariate data given the outcome (i.e., $M \perp\!\!\!\perp (A, W)|Y$). As a result, in these scenarios, the conditional OR for the effect of $A$ on $Y$ given $W$ is equal to the conditional OR given $W$ among the complete cases (Westreich, 2012). Bias of complete case estimators in scenarios 3 and 4 therefore cannot be attributed to collider stratification, despite the presence of an unobserved outcome predictor. Instead, it could result from the non-collapsibility of the odds ratio and changes in the covariate distribution brought about by narrowing the focus of inference to the complete cases (Hernán and Robins, 2017).

*Estimators*

Bagged CART was based on 100 bootstrap replicates (Lee et al., 2010). We imposed complexity constraints on the tree fitting algorithm using the `rpart` package default control settings. For boosted CART, we used 20000 iterations, a shrinkage parameter of 0.0005, and an iteration stopping rule based on the mean Kolmogorov-Smirnov test statistic (Lee et al., 2010; McCaffrey et al., 2004).

The CART methods were combined with several common approaches to handling missing data: leaving missingness information as is (i.e., subjecting incomplete data directly to the CART algorithm); complete case analysis (CCA); and multiple imputation (MI). MI was implemented with the `mice` package (version 2.46.0) using the `logreg` and `norm` options to impute missing binary and continuous variables, respectively, and otherwise default settings (Van Buuren and Groothuis-Oudshoorn, 2011). Imputation models included, apart from the variable to be imputed, all other variables, including the outcome, as untransformed main effects only. Propensity score analysis was performed within imputed datasets using the respective sets of estimated propensity scores (Penning de Vries and Groenwold, 2016) and results were combined using Rubin's (1987) rules.

In addition to using CART, as stated, we also estimated propensity scores in imputed datasets using a correctly specified propensity score model (LRc), and using a logistic model with main effects only (LRm).

Within each (multiply imputed) dataset, the ATT was estimated from a logistic model with robust variance estimation using the `survey` package (Lumley, 2014, version 3.31). We used both propensity score matching and inverse probability weighting. Matching was performed using a greedy 1:1 nearest neighbour algorithm, matching exposed ($A = 1$) to unexposed individuals ($A = 0$) (Austin, 2011a). For any given (imputed) dataset, matching was performed on the logit of the propensity score, using a calliper distance of 20% of the standard deviation of the logit propensity score estimates (Austin, 2011b). With the ATT as the estimand, IPW weights were defined as 1 for exposed subjects and PS/(1 − PS) for unexposed subjects (PS denoting the estimated propensity score). To avoid undefined weights (1/0) or logit propensity scores (logit(0)), we placed bounds on the estimated propensity scores, truncating all estimates less than 0.001 to 0.001 and setting estimates greater than 0.999 to 0.999. MI-based estimates were pooled using Rubin's rules to yield for each original dataset a single effect estimate, standard error estimate, and 90% confidence interval (90%CI).

*Performance metrics*

We evaluated the performance of the various methods through several measures: bias, estimated by the mean deviation of the estimated from the true marginal exposure-outcome effect on the log scale; empirical standard error; mean estimated standard error; mean squared error (MSE); and 90%CI coverage, estimated by the percentage of the 5000 data sets in which the 90%CI included the true exposure-outcome effect. Based on 5000 simulation runs, the Monte Carlo standard error for the true coverage probability of 0.90 is $\sqrt{}(0.90(1-0.90)/5000) \approx 0.0042$, implying that the estimated coverage probability is expected to lie with 95% probability between 0.893 and 0.907. Empirical coverage rates outside this interval provide evidence against the true coverage probabilities being equivalent to the nominal level of 0.90. The primary interest, however, was to gauge the effect of missing data on the various effect estimators. Therefore, we also compared, for each scenario, the effect estimates before and after the introduction of missing data.

### *5.3.2 Results*

In this section, we present (Table 5.2) and describe the results for IPW-based estimators only. Trends for estimators based on propensity score matching are similar and the results are presented in full in the Supplementary Material.

*Bias*

Before the introduction of missing data, baCART and bCART showed small to no bias (with absolute values ranging from 0.000 to 0.011 on the log odds ratio scale). MI+LRc performed generally well and to a similar extent as MI+baCART and MI+bCART. MI+LRm consistently underestimated the true effect when inference was based on IPW; this trend was weaker for inference based on propensity score matching (Supplementary Table 1). Among all CART-based missing data approaches considered, multiple imputation yielded the least biased estimators overall (with a maximum absolute value of bias of 0.026 versus 0.221 and 0.138 for CART-only and CCA estimators, respectively), whereas bCART deviated on average the most from the true effect after the introduction of missingness.

As expected, baCART and bCART were biased (with $-0.029$ and $-0.039$, respectively, for scenario 1 and $-0.064$ and $-0.072$ for scenario 2) under MCAR in the direction of confounding by $W_3$, whereas CCA and MI produced exposure-outcome effect estimates that were on average very close to the true effect. In scenarios 3 and 4, where missingness was outcome-dependent, bCART was biased toward the null after the introduction of missingness (with bias estimates of $-0.117$ and $0.064$ for scenario 3 and 4, respectively, where the causal log odds ratios were approximately 0.906 and $-0.926$); baCART was downwardly biased in both scenarios (with bias estimates of $-0.088$ and $-0.112$). Estimators based on CCA or MI with CART were considerably less biased. In scenarios 5 through 8, CCA estimators systematically underestimated the true effect, particularly when the effect of the exposure or the outcome on the missingness probability was large (scenarios 6 and 7, where bias estimates ranged from $-0.116$ to $-0.138$). In these scenarios (5 through 8), baCART produced estimates that were on average close to the true effect, except in scenario 6, where the effect was clearly underestimated (estimated bias $-0.050$). bCART resulted in estimates that deviated in the same direction and to a similar or greater extent from the true effect as compared with CCA estimators. Again, MI with CART resulted in estimates that were on average close to the true effect. Increasing the effect of covariate $W_1$ on the missingness probability (scenario 8 versus 5) had no evident impact on the results of any of the estimators.

### *Other performance*

As expected, discarding incomplete records (CCA) resulted in relatively large empirical standard errors. Interestingly, MI+LRc had the largest empirical standard error in most scenarios, probably as a consequence of the complexity of the fitted propensity score models. In comparing empirical and mean estimated standard errors, note that multiple imputation produced generally conservative estimates of the standard error. This is consistent with previous observations (Van Buuren, 2012). Among the CART-based estimators, the MSE was largest for CCA in nearly all scenarios. MI estimators had consistently small MSE.

Overall, the best performance in terms of MSE was attained by MI estimators, followed by baCART and bCART. Multiple imputation with CART resulted in empirical coverage rates close to or slightly higher than the nominal 90% and those of the other estimators.

### *5.3.3 Additional simulation experiment*

To investigate the estimator performances in a simpler setting, we repeated the simulation experiment of scenario 2 with the squared and interaction terms removed from the exposure allocation model of the data generating mechanism. The results, presented in Supplementary Table 2, indicate generally the same trends as previously noted. Of note, in the absence of missing data, inverse weighting based on CART showed noticeably more bias than in scenarios 1 through 8. This is probably related to CART's inherent limited ability to model smooth functions. Multiply imputing missing data followed by CART yielded approximately the same extent of bias. However, this bias appears to be partially cancelled out by the bias introduced by CART's automatic handling of missing data to the extent that CART alone performed better with than without missing data. Nonetheless, relative to the extent of bias of the respective CART algorithm before the introduction of missing data, multiple imputation with CART outperformed both CCA with CART and CART applied directly to incomplete data in terms of bias.

## 5.4 Case study

In this section, we illustrate the application of the CART-based estimators to an empirical dataset, constructed to assess the association between annual influenza vaccination and mortality risk among elderly (Groenwold et al., 2009). The dataset comprises 44418 complete records on vaccination status, mortality during the influenza epidemic period and potential confounders (age, sex, health status and prior health care and medication use). Among the 32388 vaccinated individuals 266 died, whereas 113 out of 12030 nonvaccinated individuals died (crude odds ratio 0.87, 90%CI 0.73–1.05). To control for measured confounders, propensity scores were estimated via bCART and a pseudopopulation was constructed using IPW such as to preserve the covariate distribution of the vaccination group. This resulted in an odds ratio of 0.60 (90%CI 0.49–0.73) for the marginal effect of vaccination on mortality risk among the vaccinated. Substituting bCART with baCART yielded an odds ratio of 0.65 (90% 0.53–0.81). As expected, introducing MCAR missingness into a confounder by setting

**Table 5.2**: Performance metrics of inverse probability weighting estimators in 5000 simulated datasets with and without missing data. Abbreviations: SE, standard error; MSE, mean squared error; 90%CI, 90% confidence interval; CART, classification and regression trees; baCART, bootstrap aggregated CART; bCART, boosted CART; CCA, complete case analysis; MI, multiple imputation; LRc, logistic regression with correctly specified model; LRm, logistic regression with main effects only.

| | | | Scenario | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Metric | Missing data | Method | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Bias | Without | baCART | 0.009 | 0.011 | 0.009 | 0.011 | 0.007 | 0.007 | 0.008 | 0.010 |
| | | bCART | -0.001 | 0.002 | -0.000 | 0.004 | -0.001 | -0.000 | -0.001 | 0.001 |
| | With | baCART | -0.029 | -0.064 | -0.088 | -0.112 | -0.011 | -0.050 | 0.010 | -0.004 |
| | | bCART | -0.037 | -0.072 | -0.117 | 0.064 | -0.057 | -0.221 | -0.123 | -0.053 |
| | | CCA+baCART | 0.001 | 0.001 | 0.016 | -0.023 | -0.040 | -0.138 | -0.116 | -0.032 |
| | | CCA+bCART | 0.000 | 0.006 | 0.022 | -0.010 | -0.046 | -0.136 | -0.129 | -0.035 |
| | | MI+baCART | 0.001 | 0.002 | 0.025 | 0.018 | 0.020 | 0.016 | 0.022 | 0.026 |
| | | MI+bCART | -0.008 | -0.011 | -0.024 | -0.025 | -0.010 | -0.023 | -0.008 | -0.007 |
| | | MI+LRc | 0.002 | -0.007 | -0.028 | -0.029 | -0.007 | -0.025 | -0.004 | -0.002 |
| | | MI+LRm | -0.099 | -0.094 | -0.072 | -0.075 | -0.087 | -0.083 | -0.089 | -0.082 |

Table 5.2 continued.

| | Missing | | Scenario | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Metric | data | Method | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Empirical SE | Without | baCART | 0.116 | 0.114 | 0.116 | 0.129 | 0.115 | 0.115 | 0.114 | 0.116 |
| | | bCART | 0.134 | 0.133 | 0.136 | 0.147 | 0.135 | 0.133 | 0.133 | 0.135 |
| | With | baCART | 0.118 | 0.121 | 0.126 | 0.136 | 0.119 | 0.121 | 0.119 | 0.119 |
| | | bCART | 0.132 | 0.128 | 0.125 | 0.137 | 0.131 | 0.127 | 0.149 | 0.131 |
| | | CCA+baCART | 0.141 | 0.186 | 0.189 | 0.199 | 0.147 | 0.156 | 0.143 | 0.148 |
| | | CCA+bCART | 0.158 | 0.202 | 0.211 | 0.221 | 0.165 | 0.172 | 0.154 | 0.163 |
| | | MI+baCART | 0.116 | 0.116 | 0.115 | 0.129 | 0.114 | 0.115 | 0.113 | 0.116 |
| | | MI+bCART | 0.132 | 0.130 | 0.129 | 0.140 | 0.128 | 0.126 | 0.126 | 0.128 |
| | | MI+LRc | 0.216 | 0.205 | 0.202 | 0.218 | 0.210 | 0.211 | 0.203 | 0.214 |
| | | MI+LRm | 0.125 | 0.123 | 0.125 | 0.138 | 0.122 | 0.121 | 0.124 | 0.123 |
| Mean $\widehat{\mathrm{SE}}$ | Without | baCART | 0.114 | 0.114 | 0.114 | 0.128 | 0.114 | 0.114 | 0.114 | 0.114 |
| | | bCART | 0.136 | 0.136 | 0.136 | 0.149 | 0.136 | 0.136 | 0.136 | 0.136 |
| | With | baCART | 0.115 | 0.119 | 0.118 | 0.129 | 0.117 | 0.117 | 0.119 | 0.117 |
| | | bCART | 0.134 | 0.132 | 0.131 | 0.144 | 0.134 | 0.135 | 0.153 | 0.135 |
| | | CCA+baCART | 0.139 | 0.189 | 0.190 | 0.199 | 0.148 | 0.156 | 0.145 | 0.148 |
| | | CCA+bCART | 0.160 | 0.204 | 0.211 | 0.221 | 0.166 | 0.174 | 0.160 | 0.163 |
| | | MI+baCART | 0.116 | 0.120 | 0.115 | 0.130 | 0.116 | 0.116 | 0.115 | 0.116 |
| | | MI+bCART | 0.140 | 0.143 | 0.137 | 0.150 | 0.138 | 0.138 | 0.137 | 0.138 |
| | | MI+LRc | 0.196 | 0.198 | 0.187 | 0.201 | 0.189 | 0.191 | 0.185 | 0.191 |
| | | MI+LRm | 0.131 | 0.135 | 0.128 | 0.142 | 0.128 | 0.128 | 0.128 | 0.128 |

Table 5.2 continued.

| Metric | Missing data | Method | Scenario | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| MSE | Without | baCART | 0.014 | 0.013 | 0.013 | 0.017 | 0.013 | 0.013 | 0.013 | 0.014 |
| | | bCART | 0.018 | 0.018 | 0.019 | 0.022 | 0.018 | 0.018 | 0.018 | 0.018 |
| | With | baCART | 0.015 | 0.019 | 0.024 | 0.031 | 0.014 | 0.017 | 0.014 | 0.014 |
| | | bCART | 0.019 | 0.022 | 0.029 | 0.023 | 0.020 | 0.065 | 0.037 | 0.020 |
| | | CCA+baCART | 0.020 | 0.034 | 0.036 | 0.040 | 0.023 | 0.043 | 0.034 | 0.023 |
| | | CCA+bCART | 0.025 | 0.041 | 0.045 | 0.049 | 0.029 | 0.048 | 0.040 | 0.028 |
| | | MI+baCART | 0.013 | 0.013 | 0.014 | 0.017 | 0.013 | 0.013 | 0.013 | 0.014 |
| | | MI+bCART | 0.018 | 0.017 | 0.017 | 0.020 | 0.016 | 0.016 | 0.016 | 0.016 |
| | | MI+LRc | 0.047 | 0.042 | 0.042 | 0.048 | 0.044 | 0.045 | 0.041 | 0.046 |
| | | MI+LRm | 0.025 | 0.024 | 0.021 | 0.025 | 0.023 | 0.021 | 0.023 | 0.022 |
| Empirical 90%CI coverage | Without | baCART | 0.896 | 0.901 | 0.895 | 0.897 | 0.892 | 0.899 | 0.898 | 0.890 |
| | | bCART | 0.907 | 0.909 | 0.903 | 0.904 | 0.904 | 0.909 | 0.909 | 0.907 |
| | With | baCART | 0.881 | 0.847 | 0.784 | 0.753 | 0.890 | 0.854 | 0.898 | 0.893 |
| | | bCART | 0.897 | 0.861 | 0.772 | 0.886 | 0.878 | 0.497 | 0.797 | 0.880 |
| | | CCA+baCART | 0.893 | 0.905 | 0.894 | 0.901 | 0.888 | 0.760 | 0.792 | 0.884 |
| | | CCA+bCART | 0.906 | 0.905 | 0.902 | 0.904 | 0.890 | 0.798 | 0.804 | 0.886 |
| | | MI+baCART | 0.904 | 0.914 | 0.894 | 0.897 | 0.895 | 0.900 | 0.903 | 0.896 |
| | | MI+bCART | 0.922 | 0.931 | 0.915 | 0.919 | 0.918 | 0.928 | 0.926 | 0.923 |
| | | MI+LRc | 0.911 | 0.920 | 0.909 | 0.906 | 0.908 | 0.921 | 0.919 | 0.906 |
| | | MI+LRm | 0.815 | 0.841 | 0.858 | 0.865 | 0.831 | 0.848 | 0.827 | 0.841 |

| | Missingness | | |
|---|---|---|---|
| | None | MCAR | MAR |
| Method | OR (90%CI) | OR (90%CI) | OR (90%CI) |
| baCART | 0.65 (0.53–0.81) | 0.69 (0.56–0.85) | 0.53 (0.44–0.66) |
| bCART | 0.60 (0.49–0.73) | 0.63 (0.51–0.77) | 0.79 (0.63–0.98) |
| CCA+baCART | – | 0.55 (0.41–0.73) | 0.62 (0.46–0.84) |
| CCA+bCART | – | 0.50 (0.37–0.66) | 0.56 (0.42–0.75) |
| MI+baCART | – | 0.60 (0.47–0.75) | 0.70 (0.55–0.89) |
| MI+bCART | – | 0.58 (0.47–0.72) | 0.63 (0.51–0.78) |
| LRm[†] | 0.59 (0.49–0.71) | 0.62 (0.51–0.76) | 0.70 (0.57–0.86) |

**Table 5.1:** Estimated effects of vaccination on mortality risk among the elderly in dataset with no missing data, MCAR missingness or outcome-dependent MAR missingness. Estimates are adjusted for age, sex, health status and prior health care and medication use. Abbreviations: MCAR, missing completely at random; MAR, missing at random; OR, odds ratio; 90%CI, 90% confidence interval; CART, classification and regression trees, baCART, bootstrap aggregated CART; bCART, boosted CART; CCA, complete case analysis; MI, multiple imputation; LRm, main effects logistic regression. [†]In case of (MCAR or MAR) missingness, MI was implemented before LRm.

a random 50% of subjects' number of prior general practitioner (GP) visits to missing, resulted in odds ratio estimates that were closer to the crude effect. Setting the number of GP visits to missing with probability 0.5 for all subjects who died and zero otherwise, resulted in estimates substantially closer to the null for bCART and away from the null for baCART. Thus, as in our simulations, outcome-dependent MAR missingness resulted in apparent attenuation of the exposure-outcome effect as estimated by bCART. Table 5.1 shows the results also for the complete case and multiple imputation equivalents of baCART and bCART as well as for IPW based on propensity score estimation using main effects logistic regression and with weights truncated to the interval from the 0th to the 97.5th percentile. To better handle potential violations of standard imputation model assumptions, we used a nonparametric multiple imputation strategy (option `cart` rather than `norm` in the `mice` package) to estimate the effect of vaccination on mortality risk (Doove et al., 2014). Interestingly, in the MAR setting, bCART and baCART yielded the two most extreme estimates for the effect of vaccination on mortality risk among the elderly.

## 5.5 Discussion

In this paper we examined the workings of CART based propensity score estimators in scenarios with missing covariate data. Although CART has been described as a promising approach to automatically handle missing covariate data when developing a propensity score (Setoguchi et al., 2008; Lee et al., 2010), there has been little discussion on the performance of these methods. Through analysis and simulations we showed that the application of CART for propensity score estimation can yield serious bias in estimates of exposure-outcome relations. We showed that this problem not only pertains to the situation of MAR but critically also to the situations with MCAR, which are often considered harmless when bias is concerned resulting only in larger variance of the estimator of exposure-outcome relations.

An attractive property of CART-based methods relative to standard logistic regression procedures, is perhaps not having to discard incomplete records. Indeed, in our simulations, discarding incomplete records resulted in the largest empirical standard errors. Alternatively, multiple imputation may be used to replace missing values under MCAR or MAR prior to propensity score estimation. This approach was shown to work well in our simulations. One criticism of multiple imputation in its parametric form is that it makes possibly erroneous distributional assumptions. In particular, the standard multiple imputation algorithms do not properly capture nonlinear relations like interaction effects (Cham and West, 2016). Multiple imputation algorithms that use nonparametric methods have been developed. For example, Doove et al. (2014), following Burgette and Reiter (2010), proposed CART to be incorporated as imputation method in the multiple imputation by chained equations framework. As with parametric multiple imputation, the algorithm is designed to account for the inherent variability in the data. However, while the approach of Doove et al. (2014) seems promising, there is still room for improvement. Particularly, the algorithm does not explicitly account for uncertainty about the (implicit) CART trees' model parameters. To address this, Shah et al. (2014) proposed a promising algorithm in which random forest CART is embedded in the multiple imputation by chained equations framework and imputation models are fitted to bootstrap samples. An implementation is available via the R package `CALIBERrfimpute` (Shah, 2014).

In interpreting our findings, it is important to note that we considered only a small number of scenarios. We assumed throughout that data were MCAR or MAR and that there was no unmeasured confounding (conditional exchangeability given measured confounders). As noted, there are situations conceivable in which it is not problematic to estimate the generalised propensity score. If the missingness information conveys information about a strong unmeasured confounder, estimating the generalised propensity score may allow for partial control of unmeasured confounding. On the other hand, adjusting for missingness information, e.g., through the generalised propensity score (estimated by some CART algorithms), may be problematic particularly when it is a strong proxy for the outcome, an intermediate, or a common effect of the exposure and outcome.

Our arguments for caution when using CART to estimate propensity scores in the presence of missing data are in line with the recommendation to incorporate information on the outcome in imputing missing covariate data (Penning de Vries and Groenwold, 2016; Leyrat et al., 2017; Moons et al., 2006). Since propensity score estimation is typically done without any information on the outcome (Rubin et al., 2008), any missing data imputation (e.g., with a surrogate) that is inherent to the propensity score estimation procedure will likely fail. An important feature of the propensity score matching or weighting methodology is that, in the absence of missing data, it need not make distributional assumptions about the outcome in relation to the exposure and covariates in constructing a matched or weighted dataset. In the presence of missing covariate data, omitting information on the outcome in imputing missing covariate data, however, imposes a structure on the data that likely contrasts with the true data distribution and the analysis model. This is similar to the idea of models being "uncongenial" in the sense of Meng (1994). The current study also relates to the literature on the missing indicator method, given its resemblance with the approach to handling missing data taken by the boosted CART algorithm. Like the automatic handing of missing data by the boosted CART algorithm, the missing indicator method typically results in bias (Groenwold et al., 2012).

It has been suggested to perform balance diagnostics on the matched or weighted study sample at hand (Austin, 2011a). If systematic differences persist between exposure groups following matching or weighting, this may be an indication that the propensity score estimation algorithm requires modification (Austin and Stuart, 2015). In the context of CART, one may assign greater weight to subjects at a certain covariate level in evaluating exposure homogeneity at any given node. We did not adopt an iterative approach to propensity score estimation and balance diagnostics in our simulation studies for several reasons. First, doing so would increase the computational burden of the simulations. Second, whereas CART facilitates the estimation of propensity scores that balance the entire covariate joint distribution across exposure groups, standard balance diagnostics procedures typically ignore the complex relationship between exposure and covariates. For example, when using the standardised mean difference, it is typically assumed that all variables that need to be balanced with respect to the mean are identified and included in the set over which a summary (e.g., weighted mean or maximum) standardised mean difference is calculated. The utility of the metric may be poor if important variables (e.g., higher order moments) are omitted. Other balance metrics, such as the Kolomogorov-Smirnov metric, Lévy distance, and overlapping coefficient (Belitser et al., 2011; Franklin et al., 2014; Ali et al., 2015) often fail to reflect the extent of imbalance with respect to the entire covariate joint distribution. In addition, what constitutes good balance ultimately depends on the outcome model too. Substantial imbalance may be acceptable for covariates that are weakly predictive of the outcome, while small departures from perfect balance may be problematic for covariates that are strongly predictive of the outcome.

We emphasise that our simulations were not designed to compare CART versus logistic regression as means to estimate propensity scores. Main effects logistic regression here and in previous studies demonstrated a robust performance against model misspecification in terms of bias when inference was based on propensity score matching (Setoguchi et al., 2008). This is likely attributable to the set-up of the simulations. The outcome model included homogeneous exposure-outcome effects and main effects only. Since between-exposure-group imbalances with respect to interaction terms or higher order moments of covariates need not accompany systematic differences in outcomes, it is not surprising that propensity score matching based on main effects logistic regression may perform roughly the same in terms of bias as propensity score matching based on logistic regression with correct model specification. Further studies comparing CART versus main effects logistic regression may well demonstrate more clearly the advantageous properties of CART in settings with both complex propensity score and complex outcome models.

In summary, we compared various approaches to handling missing data in estimating propensity scores via CART. While the use of machine learning in estimating propensity scores seems promising for handling complex full data structures, it unlikely represents a suitable substitute for well-established methods, such as multiple imputation, to deal with missing data.

## Appendices

### *Appendix A*

For realisations $w$ of $W$ and $a$ of $A$, let

$$\varphi(w,a) = \frac{\varphi^*(w,a)}{\mathbb{E}[\varphi^*(W,A)|A=a]}, \qquad \varphi^*(w,a) = I(a=1) + I(a=0)\frac{e(w)}{1-e(w)}.$$

We show in this subsection that weighting by $\varphi$ yields independence between covariate(s) $W$ and $A$; that is, for all $w$,

$$\varphi(w,0)\Pr(W=w|A=0) = \varphi(w,1)\Pr(W=w|A=1).$$

Also, conditional exchangeability given $W$ implies exchangeability following weighting by $\varphi$; that is, if $(Y_0,Y_1) \perp\!\!\!\perp A|W=w$ for all $w$, then

$$\begin{aligned}&\sum_w \varphi(w,0)\Pr(Y_0=y_0,Y_1=y_1,W=w|A=0)\\ &\qquad\qquad = \sum_w \varphi(w,1)\Pr(Y_0=y_0,Y_1=y_1,W=w|A=1)\end{aligned}$$

for all $y_0, y_1$.

We begin by considering $\mathbb{E}[\varphi^*(W,A)|A = a]$. It is easy to see that $\mathbb{E}[\varphi^*(W,A)|A=1] = 1$. For $a=0$, we have

$$\begin{aligned}\mathbb{E}[\varphi^*(W,A)|A=0] &= \mathbb{E}\Big[\frac{e(W)}{1-e(W)}\Big|A=0\Big]\\ &= \mathbb{E}\Big[\frac{\Pr(A=1|W)}{\Pr(A=0|W)}\Big|A=0\Big]\\ &= \sum_w \frac{\Pr(A=1|W=w)}{\Pr(A=0|W=w)}\Pr(W=w|A=0)\\ &= \sum_w \frac{\Pr(W=w|A=1)\Pr(A=1)/\Pr(W=w)}{\Pr(W=w|A=0)\Pr(A=0)/\Pr(W=w)}\Pr(W=w|A=0)\\ &= \frac{1}{\Pr(A=0)}\sum_w \Pr(W=w|A=1)\Pr(A=1)\\ &= \frac{\Pr(A=1)}{\Pr(A=0)}\end{aligned}$$

Since $\varphi(w,1) = 1$, to prove the first statement it suffices to show that $\varphi(w,0)\Pr(W=w|A=0) = \Pr(W=w|A=1)$ for all $w$. Now,

$$\varphi(w,0)\Pr(W=w|A=0)$$

$$
\begin{aligned}
&= \frac{e(w)}{1-e(w)} \frac{\Pr(A=0)}{\Pr(A=1)} \Pr(W=w|A=0) \\
&= \frac{\Pr(A=1|W=w)}{\Pr(A=0|W=w)} \frac{\Pr(A=0)}{\Pr(A=1)} \Pr(W=w|A=0) \\
&= \frac{\Pr(W=w|A=1)\Pr(A=1)}{\Pr(W=w|A=0)\Pr(A=0)} \frac{\Pr(A=0)}{\Pr(A=1)} \Pr(W=w|A=0) \\
&= \Pr(W=w|A=1),
\end{aligned}
$$

for all $w$, as desired.

To complete this proof, observe that

$$
\begin{aligned}
&\sum_w \varphi(w,0)\Pr(Y_0=y_0,Y_1=y_1,W=w|A=0) \\
&\qquad= \sum_w \frac{\Pr(A=1|W=w)}{\Pr(A=0|W=w)} \frac{\Pr(A=0)}{\Pr(A=1)} \Pr(Y_0=y_0,Y_1=y_1,W=w|A=0) \\
&\qquad= \sum_w \frac{\Pr(W=w|A=1)\Pr(A=1)/\Pr(W=w)}{\Pr(W=w|A=0)\Pr(A=0)/\Pr(W=w)} \frac{\Pr(A=0)}{\Pr(A=1)} \\
&\qquad\qquad \times \Pr(Y_0=y_0,Y_1=y_1|W=w,A=0)\Pr(W=w|A=0) \\
&\qquad= \sum_w \Pr(W=w|A=1)\Pr(Y_0=y_0,Y_1=y_1|W=w,A=0)
\end{aligned}
$$

for all $y_0, y_1$. Under conditional exchangeability given $W$, i.e., $(Y_0,Y_1) \perp\!\!\!\perp A|W$, we have $\Pr(Y_0=y_0,Y_1=y_1|W=w,A=0) = \Pr(Y_0=y_0,Y_1=y_1|W=w,A=1)$ for all $w$. Hence, $\sum_w \varphi(w,0)\Pr(Y_0=y_0,Y_1=y_1,W=w|A=0)$ becomes $\sum_w \Pr(W=w|A=1)\Pr(Y_0=y_0,Y_1=y_1|W=w,A=1)$, which is equal to $\Pr(Y_0=y_0,Y_1=y_1|A=1)$. Since $\varphi(w,1)=1$, we also have that $\sum_w \varphi(w,1)\Pr(Y_0=y_0,Y_1=y_1,W=w|A=1) = \Pr(Y_0=y_0,Y_1=y_1|A=1)$ for all $y$, which completes this proof.

### *Appendix B*

In this subsection, we give an example of a simple setting where $(Y_0,Y_1) \perp\!\!\!\perp A|e(W)$ and $W \perp\!\!\!\perp A|e^*(V)$ hold, yet $(Y_0,Y_1) \not\perp\!\!\!\perp A|e^*(V)$.

Let $W$, $A$ and $Y$ be binary mutually independent random variables and suppose that covariate missingness is MAR dependent on $Y$. Specifically, let $\Pr(W=1)=0.5$ and $\Pr(A=1|W=w)=\Pr(A=1)=0.5$ for all $w$. Further, define $Y = I(\varepsilon_Y < (1+A)/10)$, where $\varepsilon_Y \sim \mathcal{U}(0,1)$ such that $\varepsilon_Y \perp\!\!\!\perp (A,W)$. Thus, there is conditional exchangeability given $W$, so that $\Pr(Y=1|A=a,W=w) = \Pr(Y_a=1|A=a,W=w) = (1+a)/10$ for all $a, w$. Rosenbaum and Rubin (1983, Theorems 1 and 3) and Appendix A establish conditional exchangeability given $e(W)$ and exchangeability following inverse probability weighting with weights defined on the basis of $e(W)$. Now, let $\Pr(R = 0|W, A, Y, \varepsilon_Y) = 0.1 + 0.5Y$. It is easily verified that $W \perp\!\!\!\perp A|e^*(V)$. However, $e^*(V) = 4/7$ if and only if $V = *$ or, equivalently, $R = 0$. Since $R \perp\!\!\!\perp \varepsilon_Y|(A, Y)$ and $R \perp\!\!\!\perp A|Y$, for any $u \in (0, 1)$, we therefore have

$$\begin{aligned}
&\Pr(\varepsilon_Y \leq u|A = a, e^*(V) = 4/7)\\
&\qquad = \Pr(\varepsilon_Y \leq u|A = a, R = 0)\\
&\qquad = \sum_y \Pr(\varepsilon_Y \leq u|A = a, Y = y) \Pr(Y = y|A = a, R = 0)\\
&\qquad = \sum_y \Bigg\{ \Pr(\varepsilon_Y \leq u|A = a, Y = y)\\
&\qquad\qquad \times \frac{\Pr(R = 0|Y = y) \Pr(Y = y|A = a)}{\sum_{y'} \Pr(R = 0|Y = y') \Pr(Y = y'|A = a)} \Bigg\}\\
&\qquad = \sum_y \Pr(\varepsilon_Y \leq u|A = a, Y = y) \frac{(1 + 5y)[(1 + a)y + (9 - a)(1 - y)]}{15 + 5a},
\end{aligned}$$

where

$$\begin{aligned}
\Pr(\varepsilon_Y \leq u|A = a, Y = y) &= \frac{\Pr(Y = y|A = a, \varepsilon_Y \leq u) \Pr(\varepsilon_Y \leq u)}{\Pr(Y = y|A = a)}\\
&= \frac{q(y, u, a)u}{(1 + a)y/10 + (9 - a)(1 - y)/10},
\end{aligned}$$

with $q(y, u, a) = 1 - y + (-1)^{1-y}\min\{(1 + a)/10, u\}/u$. In particular, $\Pr(\varepsilon_Y \leq 0.5|A = a, e^*(V) = 4/7)$ equals $2/3$ if $a = 0$ and $3/4$ if $a = 1$. Hence, $\varepsilon_Y \not\perp\!\!\!\perp A|e^*(V)$ and, given the definitions of $Y$, $Y_0$ and $Y_1$, we have $(Y_0, Y_1) \not\perp\!\!\!\perp A|e^*(V)$.

## *Appendix C*

This subsection details an example where $(Y_0, Y_1) \perp\!\!\!\perp A|e^*(V)$, yet $(Y_0, Y_1) \not\perp\!\!\!\perp A|e(W)$.

Suppose that $W$ and that $A$ and $Y$ are all binary random variables. Further, let $(A, R)$ be marginally independent of $W$, let $A$ conditionally depend on $R$ given $W$, and let $Y$ conditionally depend on $A$ and $R$ given $W$. Specifically, let $\Pr(W = 1) = 0.5$, $\Pr(R = 0|W) = 0.1$, $\Pr(A = 1|R, W) = 2(1 + R)/10$, and $Y = I(\varepsilon_Y < 2(1 + 2R)/20)$, where $\varepsilon_Y \perp\!\!\!\perp (W, R, A)$. To see that $(Y_0, Y_1) \perp\!\!\!\perp A|e^*(V)$, first note that

$$\begin{aligned}e(w) &= \Pr(A=1|W=w)\\ &= \Pr(A=1|W=w,R=0)\Pr(R=0|W=w)\\ &\quad+\Pr(A=1|W=w,R=1)\Pr(R=1|W=w)\\ &= 0.38,\end{aligned}$$

for $w=0,1$, and that $e^*(v)$ equals $\Pr(A=1|R=0)=0.20$ if $v=*$ and $\Pr(A=1|R=1)=0.40$ otherwise. Now,

$$\begin{aligned}\Pr(Y_0=1|A=a,e^*(V)=0.20) &= \Pr(Y_0=1|A=a,e^*(V)=0.20)\\ &= \Pr(Y_0=1|A=a,R=0)\\ &= \Pr(\varepsilon_Y<2(1+2R)/20|A=a,R=0)\\ &= \Pr(\varepsilon_Y<0.10|A=a,R=0)\\ &= \Pr(\varepsilon_Y<0.10)=0.10\end{aligned}$$

for $a=0,1$. Also, $\Pr(Y_0=1|A,e^*(V)=0.40)=0.10$. Thus, $Y_0 \perp\!\!\!\perp A|e^*(V)$. Similarly, it can be shown that $(Y_0,Y_1)\perp\!\!\!\perp A|e^*(V)$. Next, observe that

$$\begin{aligned}\Pr(Y_0=1|A=a,e(W)=0.38) &= \Pr(Y_0=1|A=a)\\ &= \Pr(Y_0=1|A=a,R=0)\Pr(R=0|A=a)\\ &\quad+\Pr(Y_0=1|A=a,R=1)\Pr(R=1|A=a)\\ &= 0.10\frac{\Pr(A=a|R=0)\Pr(R=0)}{\Pr(A=a)}\\ &\quad+0.30\frac{\Pr(A=a|R=1)\Pr(R=1)}{\Pr(A=a)}\\ &= \frac{0.20^a0.80^{1-a}0.01+0.40^a0.60^{1-a}0.27}{0.38^a0.62^{1-a}},\end{aligned}$$

which is not invariant to changes in $a=0,1$. Hence, $(Y_0,Y_1)\not\perp\!\!\!\perp A|e(W)$.

## References

Albert, A. and J. Anderson (1984): "On the existence of maximum likelihood estimates in logistic regression models," *Biometrika*, 71, 1–10.

Ali, M., R. Groenwold, S. Belitser, W. Pestman, A. Hoes, K. Roes, A. de Boer, and O. Klungel (2015): "Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review," *Journal of Clinical Epidemiology*, 68, 122–131.

Austin, P. (2011a): “An introduction to propensity score methods for reducing the effects of confounding in observational studies,” *Multivariate Behavioral Research*, 46, 399–424.

Austin, P. (2011b): “Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies,” *Pharmaceutical Statistics*, 10, 150–161.

Austin, P. and E. Stuart (2015): “Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies,” *Statistics in Medicine*, 34, 3661–3679.

Belitser, S., E. Martens, W. Pestman, R. Groenwold, A. Boer, and O. Klungel (2011): “Measuring balance and model selection in propensity score methods,” *Pharmacoepidemiology and Drug Safety*, 20, 1115–1129.

Breiman, L. (1996): “Bagging predictors,” *Machine Learning*, 24, 123–140.

Breiman, L. (2001): “Random forests,” *Machine Learning*, 45, 5–32.

Burgette, L. and J. Reiter (2010): “Multiple imputation for missing data via sequential regression trees,” *American journal of epidemiology*, 172, 1070–1076.

Cham, H. and S. West (2016): “Propensity score analysis with missing data,” *Psychological Methods*, 21, 427–445.

Cole, S. and C. Frangakis (2009): “The consistency statement in causal inference: a definition or an assumption?” *Epidemiology*, 20, 3–5.

D’Agostino Jr., R. and D. Rubin (2000): “Estimating and using propensity scores with partially missing data,” *Journal of the American Statistical Association*, 95, 749–759.

Doove, L., S. van Buuren, and E. Dusseldorp (2014): “Recursive partitioning for missing data imputation in the presence of interaction effects,” *Computational Statistics & Data Analysis*, 72, 92–104.

Drake, C. (1993): “Effects of misspecification of the propensity score on estimators of treatment effect,” *Biometrics*, 49, 1231–1236.

Elith, J., J. Leathwick, and T. Hastie (2008): “A working guide to boosted regression trees,” *Journal of Animal Ecology*, 77, 802–813.

Franklin, J., J. Rassen, D. Ackermann, D. Bartels, and S. Schneeweiss (2014): “Metrics for covariate balance in cohort studies of causal effects,” *Statistics in Medicine*, 33, 1685–1699.

Groenwold, R., D. Nelson, K. Nichol, A. Hoes, and E. Hak (2009): “Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research,” *International Journal of Epidemiology*, 39, 107–117.

Groenwold, R. H., I. R. White, A. R. T. Donders, J. R. Carpenter, D. G. Altman, and K. G. Moons (2012): “Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis,” *Canadian Medical Association Journal*, 184, 1265–1269.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer, second edition.

Hernán, M. and J. Robins (2017): “Fine point 4.3: Collapsibility of the odds ratio,” in M. Hernán and J. Robins, eds., *Causal Inference*, Boca Raton: Chapman & Hall/CRC, URL `https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/`, forthcoming.

Holland, P. (1986): “Statistics in causal inference,” *Journal of the American Statistical Association*, 81, 945–960.

Holland, P. (1988): “Causal inference, path analysis, and recursive structural equations models,” *Sociological Methodology*, 18, 449–484.

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

Lesko, C., A. Buchanan, D. Westreich, J. Edwards, M. Hudgens, and S. Cole (2017): “Generalizing study results: a potential outcomes perspective,” *Epidemiology*, 28, 553–561.

Leyrat, C., S. R. Seaman, I. R. White, I. Douglas, L. Smeeth, J. Kim, M. Resche-Rigon, J. R. Carpenter, and E. J. Williamson (2017): “Propensity score analysis with partially observed covariates: How should multiple imputation be used?” *Statistical methods in medical research*, 0962280217713032.

Lumley, T. (2014): *survey: Analysis of complex survey samples (R package, version 3.31)*, Comprehensive R Archive Network, Vienna, Austria, URL `http://cran.r-project.org/web/packages/survey/index.html`.

McCaffrey, D., G. Ridgeway, and A. Morral (2004): "Propensity score estimation with boosted regression for evaluating adolescent substance abuse treatment," *Psychological Methods*, 9, 403–425.

Meng, X.-L. (1994): "Multiple-imputation inferences with uncongenial sources of input," *Statistical Science*, 538–558.

Moisen, G. (2008): "Classification and regression trees," in S. Jorgensen and B. Fath, eds., *Encyclopedia of Ecology*, volume 1, Oxford: Elsevier.

Moons, K. G., R. A. Donders, T. Stijnen, and F. E. Harrell Jr (2006): "Using the outcome for imputation of missing predictor values was preferred," *Journal of clinical epidemiology*, 59, 1092–1101.

Neyman, J., K. Iwaszkiewicz, and St. Kolodziejczyk (1935): "Statistical problems in agricultural experimentation," *Supplement to the Journal of the Royal Statistical Society*, 2, 107–180.

Pearl, J. (2009): *Causality: Models, Reasoning and Inference*, New York: Cambridge University Press.

Penning de Vries, B. and R. Groenwold (2016): "Comments on propensity score matching following multiple imputation," *Statistical Methods in Medical Research*, 25, 3066–3068.

Peters, A. and T. Hothorn (2017): *ipred: Improved Predictors (R package, version 0.9-6)*, Comprehensive R Archive Network, Vienna, Austria, URL `http://cran.r-project.org/web/packages/ipred/index.html`.

R Core Team (2016): *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `https://www.R-project.org/`.

Rai, D., B. Lee, C. Dalman, C. Newschaffer, G. Lewis, and C. Magnusson (2017): "Antidepressants during pregnancy and autism in offspring: population based cohort study," *BMJ*, 385, j2811.

Ridgeway, G. (1999): "The state of boosting," *Computing Science and Statistics*, 31, 172–181.

Ridgeway, G., D. McCaffrey, A. Morral, B. Griffin, and L. Burgette (2017): *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups (R* *package, version 1.5)*, Comprehensive R Archive Network, Vienna, Austria, URL http://cran.rproject.org/web/packages/twang/index.html.

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

Rubin, D. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66, 688–701.

Rubin, D. (1976): "Inference and missing data," *Biometrika*, 63, 581–592.

Rubin, D. (1987): *Multiple imputation for nonresponse in surveys*, New York: Wiley.

Rubin, D. B. et al. (2008): "For objective causal inference, design trumps analysis," *The Annals of Applied Statistics*, 2, 808–840.

Schafer, J. (1997): *Analysis of incomplete multivariate data*, Boca Raton: CRC Press.

Setoguchi, S., S. Schneeweiss, M. B. MA, R. Glynn, and E. Cook (2008): "Evaluating uses of data mining techniques in propensity score estimation: a simulation study," *Pharmacoepidemiology and Drug Safety*, 17, 546–555.

Shah, A. (2014): *CALIBERrfimpute: Imputation in MICE using Random Forest (R package, version 0.1-2)*, Comprehensive R Archive Network, Vienna, Austria, URL http://cran.r-project.org/web/packages/CALIBERrfimpute/index.html.

Shah, A., J. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway (2014): "Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study," *American Journal of Epidemiology*, 179, 764–774.

Stürmer, T., M. Joshi, R. Glynn, J. Avorn, K. Rothman, and S. Schneeweiss (2006): "A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. journal of clinical epidemiology," *Journal of Clinical Epidemiology*, 59, 437–e1.

Tchetgen, E. T. and T. VanderWeele (2012): "On causal inference in the presence of interference," *Statistical Methods in Medical Research*, 21, 55–75.

Therneau, T. and E. Atkinson (2017): “An introduction to recursive partitioning using the RPART routines,” *Rochester: Mayo Foundation.*

Van Buuren, S. (2012): *Flexible imputation of missing data*, Boca Raton: CRC Press.

Van Buuren, S. and K. Groothuis-Oudshoorn (2011): “mice: Multivariate imputation by chained equations in R,” *Journal of Statistical Software*, 45, 1–67.

Westreich, D. (2012): “Berkson’s bias, selection bias, and missing data,” *Epidemiology*, 23, 159–164.

Westreich, D., J. Lessler, and M. Jonsson Funk (2010): “Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression,” *Journal of clinical epidemiology*, 63, 826–833.

Wyss, R., A. Ellis, M. Brookhart, C. Girman, M. Jonsson Funk, R. LoCasale, and T. Stürmer (2014): “The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bcart, and the covariate-balancing propensity score,” *American Journal of Epidemiology*, 180, 645–655.

## Supplementary Material

**Table S5.1**: Performance metrics of propensity score matching estimators in 5000 simulated datasets with and without missing data. Abbreviations: SE, standard error; MSE, mean squared error; 90%CI, 90% confidence interval; CART, classification and regression trees; baCART, bootstrap aggregated CART; bCART, boosted CART; CCA, complete case analysis; MI, multiple imputation; LRc, logistic regression with correctly specified model; LRm, logistic regression with main effects only.

| | | | Scenario | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Metric | Missing data | Method | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Bias | Without | baCART | -0.050 | -0.048 | -0.052 | -0.046 | -0.052 | -0.052 | -0.053 | -0.051 |
| | | bCART | -0.054 | -0.053 | -0.056 | -0.047 | -0.054 | -0.055 | -0.055 | -0.054 |
| | With | baCART | -0.094 | -0.114 | -0.061 | -0.116 | -0.054 | -0.056 | -0.031 | -0.046 |
| | | bCART | -0.115 | -0.170 | -0.165 | 0.110 | -0.135 | -0.365 | -0.228 | -0.129 |
| | | CCA+baCART | -0.072 | -0.110 | -0.068 | -0.127 | -0.109 | -0.191 | -0.175 | -0.093 |
| | | CCA+bCART | -0.063 | -0.070 | -0.040 | -0.104 | -0.102 | -0.169 | -0.171 | -0.079 |
| | | MI+baCART | -0.056 | -0.054 | -0.026 | -0.031 | -0.035 | -0.033 | -0.032 | -0.030 |
| | | MI+bCART | -0.062 | -0.065 | -0.063 | -0.061 | -0.056 | -0.059 | -0.056 | -0.054 |
| | | MI+LRc | -0.010 | -0.014 | -0.006 | 0.000 | -0.006 | -0.005 | -0.004 | -0.004 |
| | | MI+LRm | -0.006 | -0.004 | -0.012 | -0.017 | -0.008 | -0.010 | -0.008 | -0.005 |

Table S5.1 continued.

| | | | Scenario | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Metric | Missing data | Method | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Empirical SE | Without | baCART | 0.133 | 0.131 | 0.130 | 0.160 | 0.131 | 0.133 | 0.132 | 0.133 |
| | | bCART | 0.146 | 0.144 | 0.147 | 0.178 | 0.144 | 0.144 | 0.144 | 0.147 |
| | With | baCART | 0.132 | 0.127 | 0.140 | 0.164 | 0.132 | 0.140 | 0.136 | 0.135 |
| | | bCART | 0.148 | 0.143 | 0.134 | 0.165 | 0.142 | 0.139 | 0.151 | 0.145 |
| | | CCA+baCART | 0.167 | 0.237 | 0.226 | 0.268 | 0.175 | 0.189 | 0.180 | 0.173 |
| | | CCA+bCART | 0.181 | 0.260 | 0.256 | 0.303 | 0.198 | 0.210 | 0.199 | 0.197 |
| | | MI+baCART | 0.127 | 0.128 | 0.124 | 0.151 | 0.123 | 0.125 | 0.124 | 0.126 |
| | | MI+bCART | 0.139 | 0.138 | 0.133 | 0.162 | 0.131 | 0.133 | 0.132 | 0.133 |
| | | MI+LRc | 0.120 | 0.118 | 0.117 | 0.142 | 0.114 | 0.117 | 0.116 | 0.116 |
| | | MI+LRm | 0.110 | 0.110 | 0.109 | 0.134 | 0.107 | 0.107 | 0.108 | 0.109 |
| Mean $\widehat{\text{SE}}$ | Without | baCART | 0.128 | 0.128 | 0.128 | 0.156 | 0.128 | 0.128 | 0.128 | 0.128 |
| | | bCART | 0.148 | 0.148 | 0.149 | 0.180 | 0.148 | 0.148 | 0.149 | 0.148 |
| | With | baCART | 0.128 | 0.126 | 0.127 | 0.153 | 0.128 | 0.127 | 0.129 | 0.127 |
| | | bCART | 0.147 | 0.146 | 0.143 | 0.172 | 0.148 | 0.147 | 0.155 | 0.148 |
| | | CCA+baCART | 0.160 | 0.237 | 0.224 | 0.262 | 0.173 | 0.182 | 0.174 | 0.173 |
| | | CCA+bCART | 0.185 | 0.265 | 0.258 | 0.299 | 0.200 | 0.212 | 0.201 | 0.200 |
| | | MI+baCART | 0.140 | 0.146 | 0.139 | 0.166 | 0.139 | 0.139 | 0.139 | 0.139 |
| | | MI+bCART | 0.163 | 0.167 | 0.161 | 0.194 | 0.161 | 0.161 | 0.161 | 0.161 |
| | | MI+LRc | 0.136 | 0.139 | 0.135 | 0.161 | 0.135 | 0.135 | 0.135 | 0.135 |
| | | MI+LRm | 0.124 | 0.127 | 0.124 | 0.150 | 0.123 | 0.124 | 0.123 | 0.123 |

Table S5.1 continued.

| | | | Scenario | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Metric | Missing data | Method | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| MSE | Without | baCART | 0.020 | 0.019 | 0.020 | 0.028 | 0.020 | 0.020 | 0.020 | 0.020 |
| | | bCART | 0.024 | 0.024 | 0.025 | 0.034 | 0.024 | 0.024 | 0.024 | 0.024 |
| | With | baCART | 0.026 | 0.029 | 0.023 | 0.041 | 0.020 | 0.023 | 0.019 | 0.020 |
| | | bCART | 0.035 | 0.049 | 0.045 | 0.039 | 0.038 | 0.152 | 0.074 | 0.038 |
| | | CCA+baCART | 0.033 | 0.068 | 0.056 | 0.088 | 0.042 | 0.072 | 0.063 | 0.039 |
| | | CCA+bCART | 0.037 | 0.073 | 0.067 | 0.103 | 0.050 | 0.073 | 0.069 | 0.045 |
| | | MI+baCART | 0.019 | 0.019 | 0.016 | 0.024 | 0.016 | 0.017 | 0.016 | 0.017 |
| | | MI+bCART | 0.023 | 0.023 | 0.022 | 0.030 | 0.020 | 0.021 | 0.020 | 0.021 |
| | | MI+LRc | 0.014 | 0.014 | 0.014 | 0.020 | 0.013 | 0.014 | 0.013 | 0.014 |
| | | MI+LRm | 0.012 | 0.012 | 0.012 | 0.018 | 0.012 | 0.012 | 0.012 | 0.012 |
| Empirical | Without | baCART | 0.860 | 0.871 | 0.867 | 0.876 | 0.866 | 0.856 | 0.858 | 0.861 |
| 90%CI | | bCART | 0.878 | 0.886 | 0.879 | 0.896 | 0.885 | 0.884 | 0.882 | 0.883 |
| coverage | With | baCART | 0.798 | 0.761 | 0.826 | 0.787 | 0.856 | 0.830 | 0.870 | 0.858 |
| | | bCART | 0.795 | 0.685 | 0.707 | 0.840 | 0.773 | 0.188 | 0.569 | 0.775 |
| | | CCA+baCART | 0.851 | 0.856 | 0.880 | 0.874 | 0.833 | 0.702 | 0.719 | 0.846 |
| | | CCA+bCART | 0.888 | 0.892 | 0.900 | 0.894 | 0.858 | 0.790 | 0.773 | 0.875 |
| | | MI+baCART | 0.901 | 0.912 | 0.928 | 0.926 | 0.926 | 0.921 | 0.921 | 0.921 |
| | | MI+bCART | 0.920 | 0.927 | 0.935 | 0.942 | 0.934 | 0.935 | 0.937 | 0.934 |
| | | MI+LRc | 0.939 | 0.947 | 0.944 | 0.942 | 0.948 | 0.943 | 0.942 | 0.946 |
| | | MI+LRm | 0.933 | 0.945 | 0.941 | 0.932 | 0.941 | 0.940 | 0.935 | 0.938 |

**Table S5.2:** Performance metrics of inverse probability weighting and matching estimators in 5000 simulated datasets of additional simulation experiment. Abbreviations: Bias dif., estimated bias after minus estimated bias before introduction missing data; Emp. SE, empirical standard error; Mean $\widehat{\text{SE}}$, mean estimated standard error; MSE, mean squared error; CART, classification and regression trees; baCART, bootstrap aggregated CART; bCART, boosted CART; CCA, complete case analysis; MI, multiple imputation; LRc, logistic regression with correctly specified model; LRm, logistic regression with main effects only.

| | | Metric | | | | | |
|---|---|---|---|---|---|---|---|
| Missing data | Method | Bias | Bias dif. | Emp. SE | Mean $\widehat{\text{SE}}$ | MSE | Coverage |
| | | *Inverse probability weighting* | | | | | |
| Without | baCART | 0.061 | ref. | 0.104 | 0.106 | 0.015 | 0.851 |
| | bCART | 0.028 | ref. | 0.119 | 0.123 | 0.015 | 0.902 |
| With | baCART | 0.008 | -0.053 | 0.113 | 0.113 | 0.013 | 0.902 |
| | bCART | -0.019 | -0.048 | 0.118 | 0.121 | 0.014 | 0.909 |
| | CCA+baCART | 0.039 | -0.023 | 0.171 | 0.178 | 0.031 | 0.906 |
| | CCA+bCART | 0.041 | 0.013 | 0.185 | 0.192 | 0.036 | 0.903 |
| | MI+baCART | 0.068 | 0.007 | 0.105 | 0.109 | 0.016 | 0.846 |
| | MI+bCART | 0.027 | -0.001 | 0.119 | 0.128 | 0.015 | 0.919 |
| | MI+LRc | 0.004 | | 0.132 | 0.141 | 0.017 | 0.927 |
| | MI+LRm | 0.005 | | 0.130 | 0.138 | 0.017 | 0.920 |

Table S5.2 continued.

| | | Metric | | | | | |
|---|---|---|---|---|---|---|---|
| Missing data | Method | Bias | Bias dif. | Emp. SE | Mean $\widehat{\text{SE}}$ | MSE | Coverage |
| | | | *Matching* | | | | |
| Without | baCART | -0.003 | ref. | 0.121 | 0.121 | 0.015 | 0.901 |
| | bCART | -0.061 | ref. | 0.133 | 0.138 | 0.021 | 0.879 |
| With | baCART | -0.045 | -0.042 | 0.124 | 0.121 | 0.017 | 0.869 |
| | bCART | -0.137 | -0.075 | 0.134 | 0.137 | 0.037 | 0.731 |
| | CCA+baCART | -0.097 | -0.094 | 0.220 | 0.222 | 0.058 | 0.871 |
| | CCA+bCART | -0.092 | -0.031 | 0.239 | 0.245 | 0.066 | 0.880 |
| | MI+baCART | 0.007 | 0.010 | 0.117 | 0.134 | 0.014 | 0.938 |
| | MI+bCART | -0.063 | -0.001 | 0.129 | 0.155 | 0.021 | 0.923 |
| | MI+LRc | 0.007 | | 0.112 | 0.131 | 0.013 | 0.946 |
| | MI+LRm | 0.007 | | 0.112 | 0.131 | 0.013 | 0.944 |

# 6

# Cautionary note: propensity score matching does not account for bias due to censoring

Bas B. L. Penning de Vries
Rolf H. H. Groenwold

*Nephrology, Dialysis and Transplantation* 2018; 33(6): 914–916

## Propensity score matching and censoring

### Abstract

This article gives a review of the limitations of propensity score matching as a tool for confounding control in the presence of censoring. Using an illustrative simulation study, we emphasize the importance of explicit adjustment for selective loss to follow-up and explain how this may be achieved.

In epidemiological research, valid causal inference is often hampered by confounding and selective loss to follow-up. Confounding is increasingly often addressed by means of propensity score (PS) matching. The analysis of a PS matched dataset closely resembles that of a randomised controlled trial (RCT); one expects that, on average, the distribution of covariates will be similar between treatment groups after propensity score matching or randomisation so that in the absence of other forms of bias systematic differences in outcomes between treatment groups can be attributed to treatment. Importantly, as is the case with RCTs (Groenwold et al., 2014). PS matching (or randomisation in the case of an RCT) typically does not account for selective loss to follow-up, and the confounder balance that was achieved through PS matching (or randomisation) may falsely reassure researchers and readers that the treatment groups under study were (and remained) comparable. The problem of selective loss to follow-up can, however, be potentially remedied by the same methods that have been proposed to address the problem in RCTs, namely inverse probability weighting, multiple imputation, or regression adjustment (Groenwold et al., 2014).

### *Two examples*

In a study on the dose-response relationship between sulfonylurea derivatives (SU) and major adverse cardiovascular events in elderly patients with type 2 diabetes, patients were censored if they switched their treatment regimen (Abdelmoneim et al., 2016). Matching on a high-dimensional PS created treatment groups (high and low dose SU) that were very similar in terms baseline characteristics, including those reflecting disease severity, comedication use, and comorbidity state. Possibly, however, those who switched treatments at any point during follow-up represent a selective subset, for example because switching occurred more often among those who used more concomitant medication. Over time, this may have distorted the balance in comedication that was initially achieved through PS matching.

Another example is a study comparing outcomes between incremental and thrice-weekly initiation of haemodialysis (Park et al., 2016). Following PS matching, the groups were similar in terms of a number of baseline characteristics including age, sex, and primary renal disease. However, approximately half of the participants were lost to follow-up at 12 months. Again, this may have induced a selection bias if the loss to follow-up affected the treatment groups differentially.

### *An illustration of the problem*

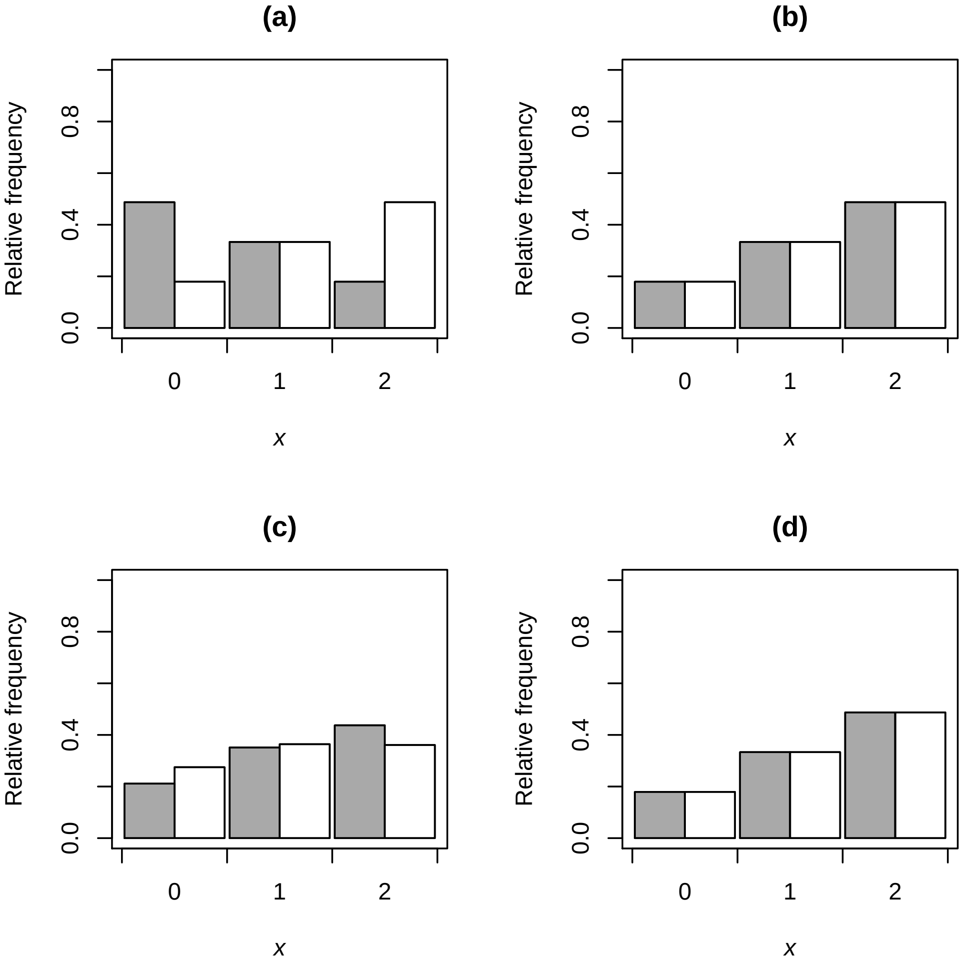
Through a small simulation study, we will illustrate the effect of ignoring selectively missing outcomes, whilst focusing on PS matching to control for confounding. Throughout, it is assumed that there is exchangeability for treatment and censoring, consistency, no model misspecification, and positivity, so that the observed covariates are sufficient to adjust for both confounding and selection bias due to loss to follow-up (Robins et al., 2000; Hernán et al., 2000; Cole and Hernán, 2008).

For this illustration, we consider a hypothetical setting representing an observational study of a binary treatment variable $T$, a binary outcome variable $Y$, and a trichotomous confounder $X$. The probability of a subject dropping out before their outcome could be assessed depends on both $T$ and $X$. Data were generated for 10,000 subjects using the mechanism detailed in the Supplementary Material. The interest lies in the marginal odds ratio (OR) of 2 for the average treatment effect on the treated (ATT). However, in this observational setting, causal inference is hampered by confounding. This motivates the use of PS matching, which typically provides an estimate of the ATT (Williamson et al., 2012). Here, PSs were estimated by a logistic regression of $T$ on $X$. We then matched treated to untreated subjects on the estimated PSs with replacement. As an alternative to PS matching to estimate the ATT, we also used inverse probability weighting, with weights of 1 and PS/(1−PS) for treated and untreated subjects, respectively. Treatment effects were estimated by applying a logistic regression to the matched or weighted pseudopopulations. We refer to these approaches as PS1 and IPW1, respectively. This procedure was repeated 1000 times. Bias was estimated on the log-odds ratio scale as the average deviation from the true log-odds ratio log 2.

The results in Table 6.1 show that both PS1 and IPW1 yielded substantial bias. The reason for this bias is apparent from Figure 6.1, which depicts the balance in the population before and after matching and/or weighting. Although PS1 and IPW1 are suited to balance confounders (Figure 6.1(a) and (b)), as subjects are lost to follow-up, the balance achieved through matching or weighting is not guaranteed to uphold in the dataset used for the analysis (Figure 6.1(c)). In fact, since the probability of dropping out depends on both $T$ and $X$, conditioning on not being lost at follow-up (i.e. performing an analysis on those subjects for whom an outcome is observed) induces an association between $X$ and $T$ (Pearl, 2009; Hernán et al., 2004), thereby biasing the relation between $T$ and $Y$ through what is formally known as collider stratification bias.

To account for selective loss to follow-up, we applied Inverse-Probability-of-

**Figure 6.1:** Balance on the confounder $X$ across treatment groups in a hypothetical setting



The untreated group is represented in grey; the treated group in white. Frequencies are relative to treatment (treated/untreated) group size; hence, equally sized bars indicate confounder balance. In the following, PS and PC denote the propensity score and the probability of censoring (being lost to follow-up) given $T$ and $X$, respectively.

Panel (a) shows the balance in the original unweighted population. Reweighting observations using weights of 1 and PS/(1 − PS) for treated and untreated subjects, respectively, results in the balance shown in (b). The same result is obtained by matching treated subjects to untreated with similar PS. Removing observations with censored outcomes from this inverse probability weighted or PS matched dataset results in imbalance (c). The balance shown in (d) is obtained by weighting the original observations with 1/(1 − PC) and PS/[(1 − PC)(1 − PS)] for treated and untreated subjects, respectively, and conditioning on noncensored observations. The same result is obtained by reweighting the PS matched dataset by 1/(1 − PC) for each subject.

Censoring-Weighting (IPCW) (Robins et al., 2000; Cole and Hernán, 2008). In this simple setting with only one point of follow-up, the IPCW weights reduce to the inverse probability of not being lost to follow-up (censored). Probabilities of censoring (PC) were estimated by logistic regression of $C$, a censoring indicator, on $T$ and $X$ applied to the original datasets. We then applied two additional estimators, PS2 and IPW2. In PS2, the matched sets obtained through PS1 were additionally weighted by 1/(1−PC) for each subject. In IPW2, the weights 1/(1−PC) and PS/[(1−PS)(1−PC)] for the treated and untreated, respectively, were applied to the original datasets, and only subjects with observed outcomes were included in the analysis. Again, treatment effects were estimated by applying a logistic regression to the matched and/or weighted pseudopopulations.

The results in Table 6.1 show that both PS2 and IPW2 yielded estimates that on average were very close to the true effect. The reason is that PS2 and IPW2

**Table 6.1:** Performance of inverse probability weighting (IPW) and PS matching estimators

| Estimator | Bias (95%CI) | OR |
|---|---|---|
| PS1 | −0.134 (−0.139, −0.129) | 1.749 |
| IPW1 | −0.135 (−0.139, −0.130) | 1.748 |
| PS2 | 0.002 (−0.003, 0.008) | 2.004 |
| IPW2 | 0.002 (−0.003, 0.007) | 2.003 |

For definitions of PS1, IPW1, PS2, and IPW2, see text. Bias was estimated by the average deviation of the estimated log-odds ratios $\hat{\beta}$ from the true effect $\beta = \log 2$ across 1000 simulated samples. 95%CI = $\bar{\hat{\beta}} - \beta \pm 1.96\sqrt{(\hat{\sigma}^2/1000)}$, where $\hat{\sigma}^2$ denotes the empirical variance of $\hat{\beta}$. OR = $\exp \bar{\hat{\beta}}$ (True OR = 2).

restore the imbalance that resulted from conditioning on not being lost to follow-up by reweighting observations such that $X$ and $T$ are no longer associated, and $X$ takes the distribution of the treated subjects (Figure 6.1(d)).

*Covariate imbalance in the absence of censoring*

It should be borne in mind that with two or more points of follow-up, covariate imbalance can develop even in the absence of censoring—specifically, that is, leaving the risk set for reasons other than sustaining the outcome of interest. Conditioning on past survival may induce an association between treatment and marginally independent covariates if past survival is a common effect of both (Hernán et al., 2004; Hernán, 2010; Aalen et al., 2015; Sjölander et al., 2016). If these covariates are also predictive of survival at a subsequent point of follow-up, this conditioning may therefore open a backdoor path, thereby inducing a selection bias. Thus, neither RCTs or PS matching or weighting analyses are guaranteed to be free of selection bias, because such selection occurs after baseline imbalances have been removed through randomisation, matching or weighting.

*Conclusion*

PS methods have gained increasing interest as means to adjust for confounding (Stürmer et al., 2006). However, as illustrated, PS matching does not account for bias due to censoring. In fact, the balance of confounders across treatment groups that was achieved by PS matching may be ruined by selective censoring. This problem can potentially be remedied by inverse probability of censoring weighting (as shown here), multiple imputation, or regression adjustment. It is important to be aware, however, that in contrast to PS matching and inverse probability weighting, the estimand of conventional multivariable regression analysis is not typically a marginal effect such as the ATT. Also, our simulations were done under the assumption that the censoring mechanism was independent of the outcome. Importantly, neither of the above methods is suited to solve the problem of censored data when the missingness depends on unobserved variables that are predictive of the outcome or on the outcome itself. It is only when the missingness can be explained by observed data, such as in our illustration, that such biases may be adequately addressed by one of the above methods. If loss to follow-up is a completely random process, the confounder balance that was achieved by PS matching is expected to be preserved and conventional analysis on those for whom an outcome was observed will still be appropriate.

## References

Aalen, O. O., R. J. Cook, and K. Røysland (2015): "Does cox analysis of a randomized survival study yield a causal treatment effect?" *Lifetime data analysis*, 21, 579–593.

Abdelmoneim, A. S., D. T. Eurich, A. Senthilselvan, W. Qiu, and S. H. Simpson (2016): "Dose-response relationship between sulfonylureas and major adverse cardiovascular events in elderly patients with type 2 diabetes," *Pharmacoepidemiology and drug safety*.

Cole, S. R. and M. A. Hernán (2008): "Constructing inverse probability weights for marginal structural models," *American journal of epidemiology*, 168, 656–664.

Groenwold, R. H., K. G. Moons, and J. P. Vandenbroucke (2014): "Randomized trials with missing outcome data: how to analyze and what to report," *Canadian Medical Association Journal*, 186, 1153–1157.

Hernán, M. A. (2010): "The hazards of hazard ratios," *Epidemiology (Cambridge, Mass.)*, 21, 13.

Hernán, M. Á., B. Brumback, and J. M. Robins (2000): "Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men," *Epidemiology*, 11, 561–570.

Hernán, M. A., S. Hernández-Díaz, and J. M. Robins (2004): "A structural approach to selection bias," *Epidemiology*, 15, 615–625.

Park, J. I., J. T. Park, Y.-L. Kim, S.-W. Kang, C. W. Yang, N.-H. Kim, Y. K. Oh, C. S. Lim, Y. S. Kim, and J. P. Lee (2016): "Comparison of outcomes between the incremental and thrice-weekly initiation of hemodialysis: a propensity-matched study of a prospective cohort in korea," *Nephrology Dialysis Transplantation*, gfw332.

Pearl, J. (2009): *Causality: models, reasoning, and inference*, Cambridge university press.

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

Sjölander, A., E. Dahlqwist, and J. Zetterqvist (2016): "A note on the noncollapsibility of rate differences and rate ratios," *Epidemiology*, 27, 356–359.

Stürmer, T., M. Joshi, R. J. Glynn, J. Avorn, K. J. Rothman, and S. Schneeweiss (2006): "A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods," *Journal of clinical epidemiology*, 59, 437–e1.

Williamson, E., R. Morley, A. Lucas, and J. Carpenter (2012): "Propensity scores: from naive enthusiasm to intuitive understanding," *Statistical methods in medical research*, 21, 273–293.

## Supplementary Material

In our hypothetical setting, the mechanism for generating data is defined by sequentially sampling for each subject (independently) from the following distributions. Covariate $X$ takes values 0, 1, and 2 only, each with probability 1/3. $T|X = x$ has the Bernoulli distribution with probability $\Pr(T = 1|x) = \text{expit}\{-1 + x\}$, were $\Pr(T = 1|x)$ is shorthand notation for $\Pr(T = 1|X = x)$. $C|x, t$ has the Bernoulli distribution with $\Pr(C = 1|x, t) = \text{expit}\{-1.5 + 0.5x + 2t\}$. Finally, $Y|x, t, c$ has the Bernoulli distribution with probability $\Pr(Y = 1|x, t, c) = \text{expit}\{-1 + x + 0.789t\}$. Potential outcomes $Y_{\check{t},\check{c}}$ under the combined treatment and censoring state $(\check{t}, \check{c})$ are distributed such that $\Pr(Y_{\check{t},\check{c}} = 1|x, t, c) = \Pr(Y = 1|x, t, c) = \text{expit}\{-1 + x + 0.789\check{t}\}$. By the law of total probability, $\Pr(Y_{\check{t},\check{c}} = 1|t) = \sum_{c=0}^{t} \sum_{x=0}^{2} \Pr(Y_{\check{t},\check{c}} = 1|x, t, c) \Pr(C = c|x, t) \Pr(X = x|t)$, where, by Bayes' theorem, $\Pr(X = x|t) = \Pr(T = t|x) \Pr(X = x)/ \sum_{x=0}^{2}[\Pr(T = t|x) \Pr(X = x)]$. The interest lies in the marginal odds ratio $\theta$ for the treatment effect on the treated, if contrary to fact all subjects had remained uncensored; $\theta = \text{Odds}(\Pr(Y_{1,0} = 1|T = 1))/\text{Odds}(\Pr(Y_{0,0} = 1|T = 1))$, where $\text{Odds}(p) = p/(1 - p)$. It follows that $\theta \approx 2$.

# 7

# Bias of time-varying exposure effects due to time-varying covariate measurement strategies

Bas B. L. Penning de Vries
Rolf H. H. Groenwold

*Pharmacoepidemiology and Drug Safety* 2021; 1–6

## Abstract

*Purpose.* In studies of effects of time-varying drug exposures, adequate adjustment for time-varying covariates is often necessary to properly control for confounding. However, the granularity of the available covariate data may not be sufficiently fine, for example when covariates are measured for participants only when their exposure levels change. *Methods.* To illustrate the impact of choices regarding the frequency of measuring time-varying covariates, we simulated data for a large target trial and for large observational studies, varying in covariate measurement design. Covariates were measured never, on a fixed-interval basis, or each time the exposure level switched. For the analysis, it was assumed that covariates remain constant in periods of no measurement. Cumulative survival probabilities for continuous exposure and non-exposure were estimated using inverse probability weighting to adjust for time-varying confounding, with special emphasis on the difference between five-year event risks. *Results.* With monthly covariate measurements, estimates based on observational data coincided with trial-based estimates, with five-year risk differences being zero. Without measurement of baseline or post-baseline covariates, this risk difference was estimated to be 49% based on the available observational data. With measurements on a fixed-interval basis only, five-year risk differences deviated from the null, to 29% for six-monthly measurements, and with magnitude increasing up to 35% as the interval length increased. Risk difference estimates diverged from the null to as low as −18% when covariates were measured depending on exposure level switching. *Conclusion.* Our simulations highlight the need for careful consideration of time-varying covariates in designing studies on time-varying exposures. We caution against implementing designs with long intervals between measurements. The maximum length required will depend on the rates at which treatments and covariates change, with higher rates requiring shorter measurement intervals.

## 7.1 Introduction

In many pharmacoepidemiologic studies, the use of the drugs that are investigated may change over time. In case of such time-varying exposures, the exposure effect can be defined in different ways. For example, one could contrast initiating drug treatment at a particular point in time (irrespective of whether the use is continued) with not initiating, or continuous drug use with continuous non-use. While analyses of point interventions (e.g., a single-dose vaccination) require adjustment for confounding at baseline only, for analyses of a time-varying exposure, information on time-varying covariates might be required to mitigate bias due to time-varying confounding. However, the granularity of the available information about the time-varying covariates may not be sufficiently fine to adequately control for confounding.

One special case of where this issue may arise is where researchers choose to measure covariates for study subjects only when their exposure levels have changed since the last measurement. If exposure levels do not change, covariate levels are (implicitly) assumed to remain constant, which is an implementation of a method generally known as last-observation-carried-forward (LOCF). The accurateness of the observed covariate data may then depend on the observed exposure history. In studies of antidepressant use and the risk of hip fracture, for example, comorbidities and use of co-medication were assessed only at baseline and whenever patients switched exposure level or after every six months in the absence of switching (Ali et al., 2016; Souverein et al., 2016).

In this paper, we investigate the impact of various covariate measurement designs on the estimation of time-varying exposure effects in observational studies with time-varying confounding. We illustrate, by way of simulation, the potential for bias of inverse-probability-weighting (IPW) estimators under static designs of fixed-interval covariate measurement and under dynamic designs with covariates being measured depending on the observed exposure history. IPW estimators are considered as these are increasingly used for estimating causal effects of time-varying exposures, can accommodate exposure-covariate feedback (Hernán and Robins, 2020), and readily allow for 'adjusted' survival curves to be created (Cole and Hernán, 2004).

## 7.2 Methods

We first simulated data for a hypothetical study, the 'target trial', which if implemented on theoretical population of interest would readily allow us to identify the exposure effect of interest (Hernán and Robins, 2016). In practice, it is not always possible to implement a target trial, but we use it here as a means to clarify the exposure effect of interest and we simulate from it to give a reference against which to compare results from analyses that are based on simulated data for observational studies. We considered multiple observational studies, each with the same data-generating mechanism but with different covariate measurement designs to evaluate their impact. Having simulated data, we then estimated the survival curves for the period of five years, using a weighting approach (described below) that was designed to keep treatment arms comparable throughout follow-up in terms of measured covariates. For each of the trial and observational studies, we first generated data on a single sample of $n = 150\,000$ individuals, which is sufficiently large to allow us to ignore sampling variability and regard differences between the survival curves as measures of the impact of the measurement designs on the large sample bias of the IPW estimators. The results corresponding to this single simulation run are described in detail below. In the online supplementary material, we summarise the results of 5000 independent simulation runs for sample sizes 150 000, 10 000, 1000, and 100. R code for the simulations is also provided as online supplementary material.

### *Set-up*

The target trial has the following key design elements: (1) study participants (subjects who satisfy the eligibility criteria) are randomised at a well-defined baseline time point $t_0$ to either being issued a drug prescription ($A_0 = 1$)—say, a prescription for a daily dose of some antidepressant drug for the next one-month period—or to not being issued the prescription ($A_0 = 0$) at $t_0$; (2) participants are then followed over time until the occurrence of an event (e.g., the first hip fracture or death if the subject dies without having sustained a hip fracture during follow-up) or the administrative study end, whichever comes first; (3) provided event-free survival is long enough, study participants in the ($A_0 = 1$)-group are issued a further prescription after every month since $t_0$ and those in the ($A_0 = 0$)-group do not receive a prescription during follow-up. For a given subject, we define $A_k$ to be the indicator variable that takes the value of 1 if the subject is on a one-month prescription on month $k$; $A_k = 0$ otherwise. We further define $Y$ to be the amount of follow-up time between baseline and the subject's (first) event and let $Y_k$ be that part of $Y$ that relates to month $k$. We stipulate that study participants are event-free at the start of the study and that subjects do not get lost to follow-up before the administrative study end, which we stipulated to be five years (or $K = 60$ months) post-baseline.

The observational studies differ from the target trial in the following ways only: (1) the decision to allocate a subject to $A_0 = 1$ versus $A_0 = 0$ is not made by randomisation; (2) the decisions to renew prescriptions for subjects in the $(A_0 = 1)$-group or to never issue a prescription throughout the follow-up period for those in the $(A_0 = 0)$-group are not determined by their baseline allocations $A_0$. Rather, for month $k = 0, 1, ...$, the decision to set exposure $A_k$ to 0 or 1 is based only on past exposure history $(A_j : j < k)$ and certain binary covariates $L_k$. In this observational setting, subjects can switch at the start of each month between exposure levels 'being on prescription' (or 'exposed') versus 'not being on prescription' (or 'not exposed'). In variations on this setting, covariate data were measured according to one of the following measurement designs: (1) covariates were not measured at all, thus precluding any adjustment for confounding and effectively forcing us to implement a 'crude' estimator; (2) covariates were measured on a monthly basis, which is sufficient for identification of our target quantity; (3) covariates were measured on a six-monthly basis starting at baseline; (4) covariates were measured when the respective subject's exposure level switched; (5) covariates were measured with an exposure level switch and at a six-monthly basis in the absence of exposure level switching. We also considered variations on designs (3) and (5) where, instead of six months, the fixed measurement interval have a length of 2, 3, 9, 12, ..., or 60 months. Where design (3) means that measurement times are known before the start of follow-up, designs (4) and (5) are dynamic in the sense that whether or not a subject's covariate level is measured depends on the subject's time-varying variables.

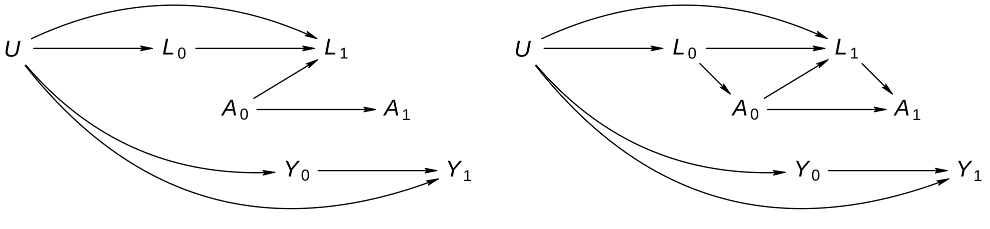


**Figure 7.1:** Directed acyclic graphs representing the data-generating mechanism for the first two months of the target trial (left) and observational study (right). Here, $U$ represents a unmeasured common cause of the measured covariates $L_0, L_1$ and outcome variables $Y_0, Y_1$. The absence of directed paths from exposure variables to outcome variables reflects the absence of a causal exposure-outcome effect.

### *Data-generating mechanism*

To simulate longitudinal data for a setting with time-varying confounding we used a variation on the approaches described by Havercroft and Didelez (2012) and Young and Tchetgen Tchetgen (2014). The data-generating mechanisms for the target trial and observational studies are described in the Appendix and produce data that are consistent with the directed acyclic graphs (DAGs) of Figure 7.1. In the trial setting (left panel of Figure 7.1), the absence of arrows going into the exposure variables reflects the absence of (time-varying) confounding. In the target trial, post-baseline exposures are fully determined by the baseline level of exposure, which takes the value of 1 for half of subjects (i.e., exposure status does not change over time). In the observational study, however, approximately 40% of subjects will have switched exposure level by the end of follow-up in each of the arms that are defined by baseline exposure level.

### *Defining and estimating the exposure effect*

We define the exposure effect of interest as a contrast between continuous exposure ($A_j = 1$ for $j = 0, 1, ...$) versus continuous non-exposure ($A_j = 0$ for $j = 0, 1, ...$). In particular, we suppose that the interest lies with a contrast between the five-year event-free survival probabilities that we would observe had everyone received continuous exposure versus continuous non-exposure; i.e., a contrast that is identified in the target trial as

$$\Pr(Y \geq 60 | A_0 = 1) \quad \text{versus} \quad \Pr(Y \geq 60 | A_0 = 0).$$

As indicated by the absence of a directed path of arrows from the exposure variables to the outcome variables in the DAG for the target trial, the difference between these two survival probabilities is zero.

To account for time-varying confounding in the observational studies, we implemented IPW by applying a crude (Kaplan-Meier) estimator to an artificial data set where, given any time during follow-up, a subject received a weight of zero if the subject had experienced an exposure level switch by that time and otherwise a weight equal to the reciprocal of the product of the estimated probabilities of their observed exposure levels until that time given the respective measured exposure and covariate histories. That is, for $a = 0, 1$, a subject's weight for month $k$ was

$$W_k = \prod_{j=0}^{k} \frac{1}{\Pr(A_j = a | Y \geq j, A_0 = ... = A_{j-1} = a, L_0, ..., L_j)}$$

if the subject received exposure level $a$ in months 0 through $k$ (i.e., $A_0 = ... = A_k = a$). Subjects were censored (i.e., received a weight of zero) from the time at which they switched to another exposure level. Apart from the covariate measurement design, the validity of the approach also rests on the correct specification of the model for the conditional treatment probabilities. To ensure correct specification for the reference measurement design (1), we assumed that the exposure $A_k$ given survival and past exposure and covariate levels was Bernoulli distributed with mean equal to

$$\begin{aligned}\Pr(A_k = 1|Y \geq k, A_0, ..., A_{k-1}, L_0, ..., L_k)\\ = \frac{\exp[\alpha_0 + \alpha_1 I(k = 0) + \alpha_2 A_{k-1} + \alpha_3 L_k]}{1 + \exp[\alpha_0 + \alpha_1 I(k = 0) + \alpha_2 A_{k-1} + \alpha_3 L_k]}\end{aligned}$$

for some $\alpha_0, \alpha_1, \alpha_2, \alpha_3$, which were estimated by a pooled logistic regression under this model. Throughout, variables that were unobserved by measurement design were handled with LOCF.

## 7.3 Results

Figure 7.2 shows the estimated survival curves for the ‘always treat’ and ‘never treat’ protocols. Consistent with the absence of a directed path from the exposure variables to the outcome variables in the DAGs of Figure 7.1, the trial-based estimates of the survival curves overlap (Figure 7.2, panel A). Where we observed a five-year event risk of 31% in both arms of the target trial, in the observational setting, we observed a risk of 64% and 15% in those who do and those who do not receive a treatment prescription at baseline, respectively, giving a risk difference of 49% (panel B). With monthly covariate measurement, IPW resulted in survival curves that virtually coincide with those of the trial (panel C), for which we found a risk difference of zero. Six-monthly measurements (panel D), however, brought the curves closer to those of the no measurement setting (panel B), i.e., in the ‘direction of confounding’. The five-year risks with six-monthly measurements were estimated to be to 50% and 21%, respectively, giving a risk-difference of 29%. In Figure 7.3, panel A, it is shown that the estimated risk differences at two and five years increase with the interval measurement length, until they reach a plateau of approximately 20% and 35%, respectively. When the interval length was set equal to the maximum follow-up duration (60 months), only baseline covariates were measured, which resulted in an estimated five-year risk difference that was approximately 15 percent points closer to the target than that of no covariate measurement at all (Figure 7.2, panel B). When we

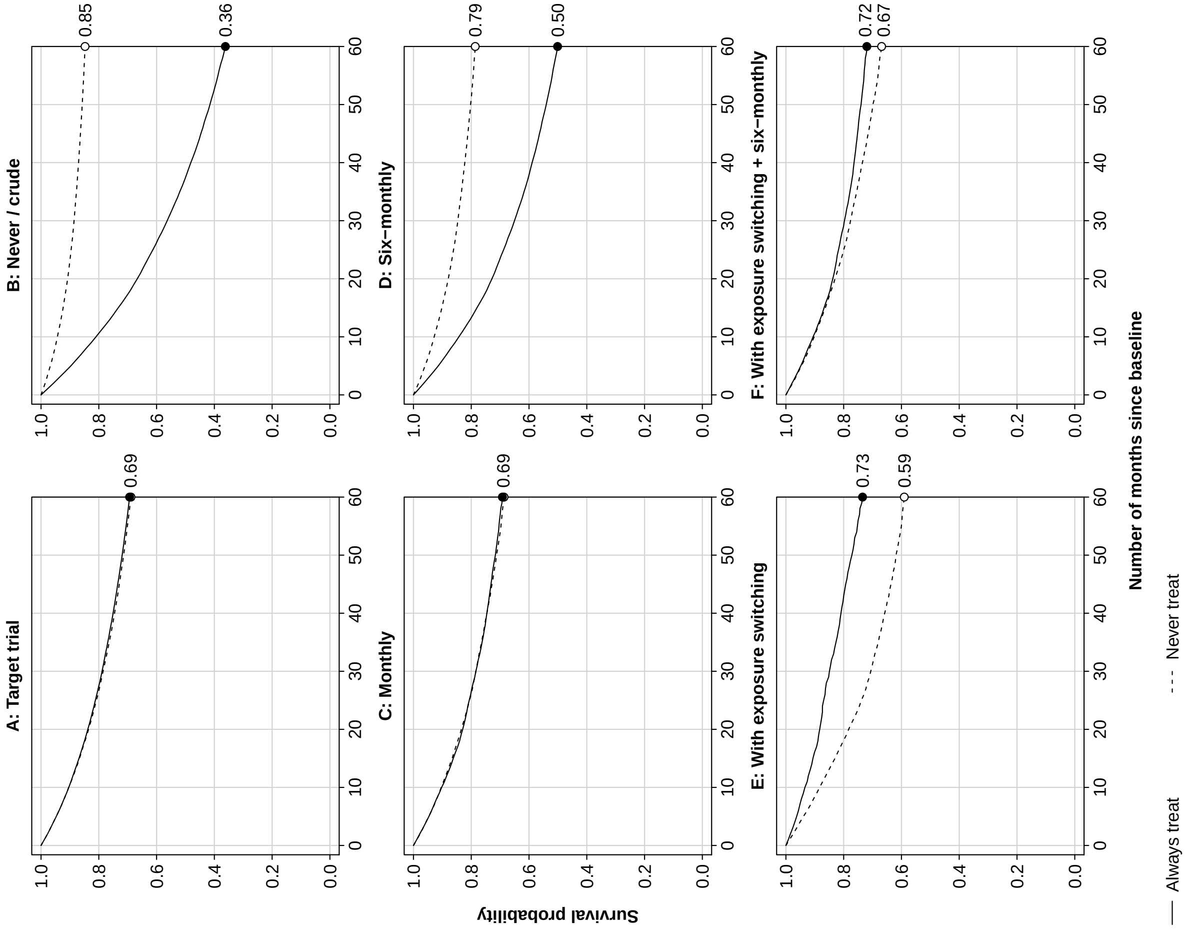


**Figure 7.2:** Estimated event-free survival curves for 'always treat' and 'never treat' protocols based on target trial (panel A) and observational study (B through F) with varying covariate measurement designs: no covariate measurement (B), continuous to monthly covariate measurement (C), six-monthly covariate measurement (D), covariate measurement only with covariate level switching (E), and with exposure switching and six-monthly in periods without switching (F).

implemented measurement design (4), the estimated 5-year risk difference flipped to the other side of the null, −14% (panel E), with five-year risks estimated to be 27% and 41% for the 'always treat' and 'never treat' protocols, respectively. For design (5), we observed a 5-year risk difference of −5%, somewhere between the results of design (3) and (4) (panel F). With increasingly large measurement intervals within periods of no switching, the estimated two-year risk difference steadily decreased to approximately −15% (Figure 7.3, panel B). The estimated five-year risk was also −15% with 60 months between measurements in periods of no switching, equal to the observed risk of design (4), as expected. However, it was lowest, approximately −18%, with an interval length of around 30 months.

The bias estimates of the survival curves and 5-year risk differences that were derived by averaging across 5000 independent samples of sizes 150 000, 10 000 and 1000 are nearly identical to the corresponding estimates described above and given in Figures 7.2 and 7.3 (cf. online supplementary material). For sample size 100, however, we observed substantial (small sample) bias for all measurement designs, even in the reference observational setting with full/monthly covariate measurement.

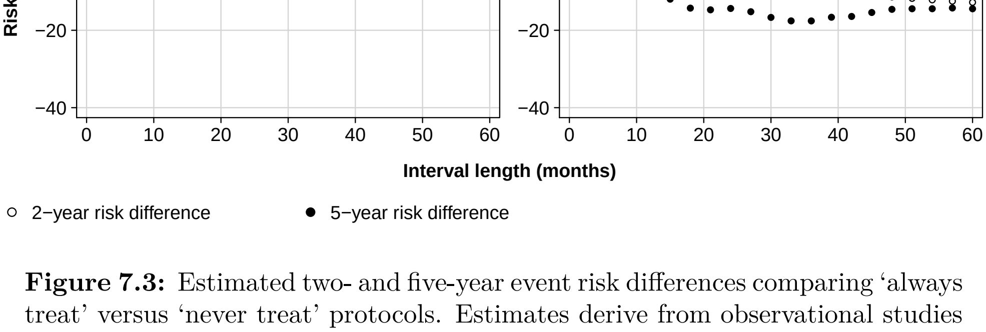


**Figure 7.3:** Estimated two- and five-year event risk differences comparing 'always treat' versus 'never treat' protocols. Estimates derive from observational studies with varying covariate measurement designs. Panel A gives the estimates for fixed-interval measurement; panel B gives the estimates for covariate measurement with exposure switching and with fixed-length intervals in periods without switching.

## 7.4 Discussion

We used simulation to study and illustrate the potential for bias due to measurement design choices in the estimation of the effects of time-varying exposures. The potential for bias in settings with static or fixed-interval covariate measurement designs has recently been illustrated already (Young et al., 2019). We additionally showed that bias might arise in settings where decisions to measure are driven by observed values of the time-varying exposure.

As expected, in our simulations, fixed-interval measurement resulted in bias in the direction of confounding, bias that is attributable to residual confounding. Interestingly, we found bias in the opposite direction when we implemented measurement designs where covariates were measured preferentially with exposure level switches. Together with LOCF, these measurement designs introduced a form of differential misclassification, which may result in bias even in the absence of confounding (Webster-Clark et al., 2020). Researchers familiar with DAGs might be alerted by the presence of colliders in the DAG that encodes part of the misclassification mechanism. For example, on the DAG of the right panel of Figure 7.1, the differential misclassification of $L_1$ can be represented by adding a measured version of $L_1$ with incoming arrows from $L_0$, $L_1$, $A_0$ and $A_1$. The measured variable can then be seen to be a collider on the path from $A_1$ to $Y_1$ via $L_1$ and $U$. By conditioning on the collider (and not the unmeasured variable $L_1$ or $U$), the path is opened, potentially leading to collider-stratification bias (Hernán and Robins, 2020).

In addition to adequate measurement of the time-varying covariates, the validity of IPW rests on the correct specification of the model for the distribution of the treatment variables given survival and past covariate and exposure levels. It is possible that the biases that we observed are partly due to model misspecification.

We considered a specific and relatively simple setting with a single, binary covariate, no censoring before the administrative study end and an interest in static rather than dynamic treatment strategies. These features are not required for valid inference with IPW (Hernán and Robins, 2020). However, the magnitude and direction of bias in other settings may differ from those observed in the current study. We stress that the bias that was observed in our simulation does not depend critically on the choice of IPW as a means to control for time-varying confounding. The choices regarding the frequency of covariate measurements will likely also affect the properties other methods, including the commonly applied Cox' regression analysis with time-varying covariates. The extent to which such choices impact a particular study are obviously context-specific. For example, it will likely depend on the rate at which subjects cross over between treatment arms as well as on the extent to which covariates are subject to change over time.

In conclusion, our simulations highlight the need for adequate measurement of time-varying covariates in observational studies on the effects of time-varying exposures. Researchers should consider differential covariate misclassification as a possible source of bias when designing covariate measurement strategies (Webster-Clark et al., 2020). Whether or not covariates are measured with every exposure level switch, we caution against implementing measurement designs with long intervals between measurements, particularly when the impact of the design choices are poorly understood. The maximum interval length that is sufficient to yield negligible bias will depend on the rates at which treatments and covariates can change (Young et al., 2019), with higher rates requiring shorter measurement intervals.

## References

Ali, M. S., R. H. Groenwold, S. V. Belitser, P. C. Souverein, E. Martín, N. M. Gatto, C. Huerta, H. Gardarsdottir, K. C. Roes, A. W. Hoes, et al. (2016): "Methodological comparison of marginal structural model, time-varying cox regression, and propensity score methods: the example of antidepressant use and the risk of hip fracture," *Pharmacoepidemiology and Drug Safety*, 25, 114–121.

Cole, S. R. and M. A. Hernán (2004): "Adjusted survival curves with inverse probability weights," *Computer methods and programs in biomedicine*, 75, 45–49.

Havercroft, W. and V. Didelez (2012): "Simulating from marginal structural models with time-dependent confounding," *Statistics in medicine*, 31, 4190–4206.

Hernán, M. and J. Robins (2020): *Causal Inference: What If*, Boca Raton: Chapman & Hall/CRC.

Hernán, M. A. and J. M. Robins (2016): "Using big data to emulate a target trial when a randomized trial is not available," *American journal of epidemiology*, 183, 758–764.

Souverein, P. C., V. Abbing-Karahagopian, E. Martin, C. Huerta, F. de Abajo, H. G. Leufkens, G. Candore, Y. Alvarez, J. Slattery, M. Miret, et al.

(2016): "Understanding inconsistency in the results from observational pharmacoepidemiological studies: the case of antidepressant use and risk of hip/femur fractures," *Pharmacoepidemiology and Drug Safety*, 25, 88–102.

Webster-Clark, M., M. Jonsson Funk, and T. Stürmer (2020): "Single-arm trials with external comparators and confounder misclassification: How adjustment can fail," *Medical Care*, 58, 1116–1121.

Young, J. G. and E. J. Tchetgen Tchetgen (2014): "Simulation from a known Cox MSM using standard parametric models for the g-formula," *Statistics in medicine*, 33, 1001–1014.

Young, J. G., R. Vatsa, E. J. Murray, and M. A. Hernan (2019): "Interval-cohort designs and bias in the estimation of per-protocol effects: a simulation study," *Trials*, 20, 552.

## Supplementary Material

**Table S7.1:** Summary of estimated 5-year (always-versus-never-exposed) risk differences over 5000 simulation runs for sample sizes 150 000, 10 000, 1000, and 100. (Continued on next page.)

| Study/measurement design[†] | Mean estimate (95% CI)[‡] | Empirical variance | Mean squared error |
|---|---|---|---|
| | *Sample size: 150 000* | | |
| A: Target trial | -0.000 (-0.000, 0.000) | 0.000 | 0.000 |
| B: Observational study 1 | 0.485 (0.485, 0.485) | 0.000 | 0.235 |
| C: Observational study 2 | -0.000 (-0.000, 0.000) | 0.000 | 0.000 |
| D: Observational study 3 | 0.286 (0.286, 0.286) | 0.000 | 0.082 |
| E: Observational study 4 | -0.134 (-0.134, -0.134) | 0.000 | 0.018 |
| F: Observational study 5 | -0.044 (-0.044, -0.043) | 0.000 | 0.002 |
| | *Sample size: 10 000* | | |
| A: Target trial | 0.000 (-0.000, 0.000) | 0.000 | 0.000 |
| B: Observational study 1 | 0.485 (0.485, 0.486) | 0.000 | 0.236 |
| C: Observational study 2 | 0.000 (-0.000, 0.001) | 0.001 | 0.001 |
| D: Observational study 3 | 0.286 (0.286, 0.287) | 0.000 | 0.082 |
| E: Observational study 4 | -0.135 (-0.136, -0.134) | 0.002 | 0.021 |
| F: Observational study 5 | -0.043 (-0.044, -0.042) | 0.001 | 0.003 |

[†] The target trial and observational studies are described in the main text. Observational studies 1 through 5 differ in covariate measurement design: in observational study 1 (B), covariates were never measured; in study 2 (C), covariates were measured on a monthly basis; in study 3 (D), covariates were measured on a six-monthly basis starting at baseline; in study 4 (E), covariates were measured when the respective subject's exposure level switched; ...

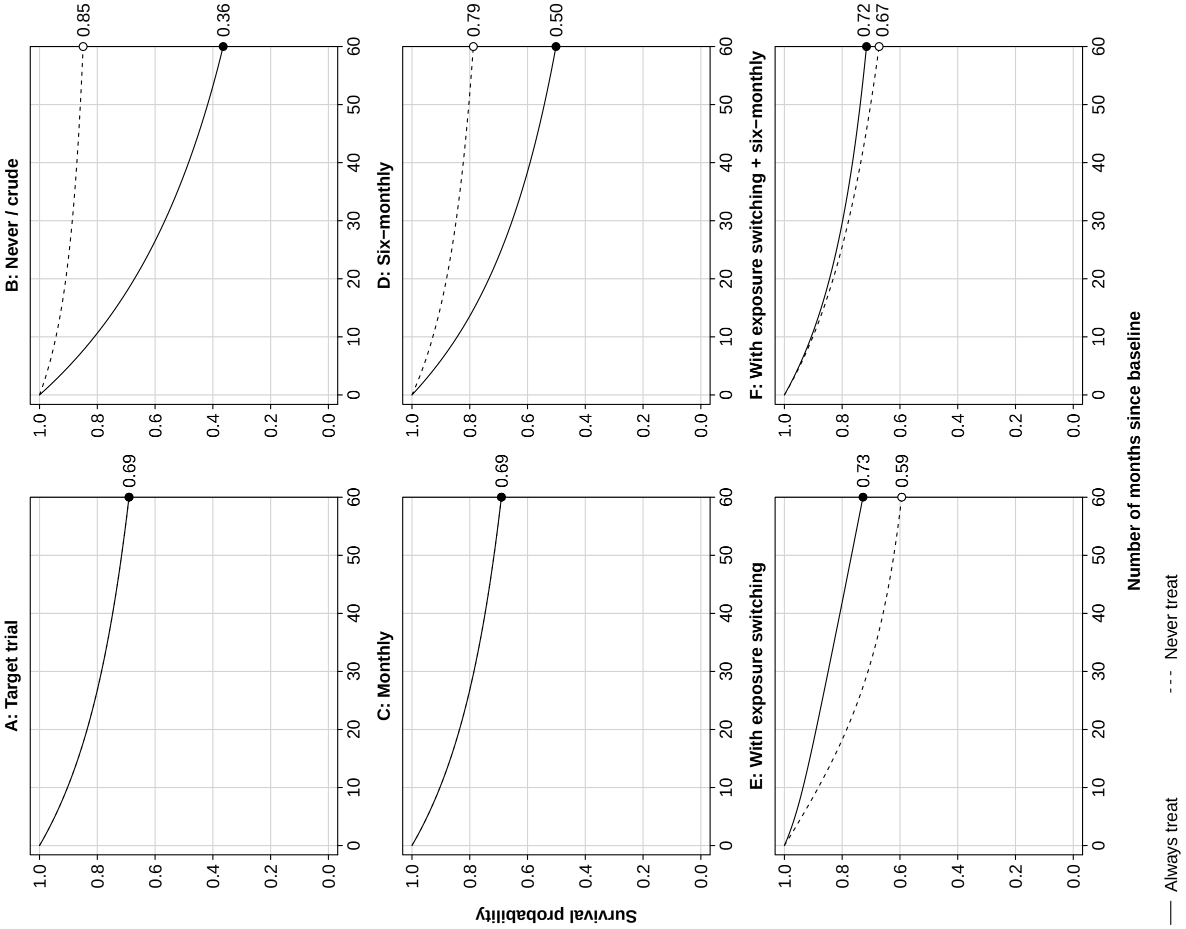


**Figure S7.1:** Mean estimated event-free survival probabilities across 5000 samples of size 150 000 based on target trial (panel A) and observational study (B through F) with varying covariate measurement designs: no covariate measurement (B), continuous to monthly covariate measurement (C), six-monthly covariate measurement (D), covariate measurement only with covariate level switching (E), and with exposure switching and six-monthly in periods without switching (F).

Table S7.1 continued.

| Study/measurement design[†] | Mean estimate (95% CI)[‡] | Empirical variance | Mean squared error |
|---|---|---|---|
| | *Sample size: 1000* | | |
| A: Target trial | 0.000 (-0.001, 0.001) | 0.001 | 0.001 |
| B: Observational study 1 | 0.486 (0.485, 0.487) | 0.002 | 0.238 |
| C: Observational study 2 | 0.014 (0.012, 0.016) | 0.007 | 0.007 |
| D: Observational study 3 | 0.291 (0.289, 0.292) | 0.003 | 0.087 |
| E: Observational study 4 | -0.132 (-0.136, -0.128) | 0.026 | 0.044 |
| F: Observational study 5 | -0.028 (-0.031, -0.025) | 0.009 | 0.009 |
| | *Sample size: 100* | | |
| A: Target trial | -0.000 (-0.003, 0.003) | 0.009 | 0.009 |
| B: Observational study 1 | 0.484 (0.481, 0.488) | 0.016 | 0.251 |
| C: Observational study 2 | 0.117 (0.110, 0.123) | 0.060 | 0.073 |
| D: Observational study 3 | 0.320 (0.315, 0.324) | 0.026 | 0.129 |
| E: Observational study 4 | -0.004 (-0.014, 0.007) | 0.154 | 0.154 |
| F: Observational study 5 | 0.091 (0.084, 0.099) | 0.072 | 0.080 |

... in study 5 (F), covariates were measured with an exposure level switch and at a six-monthly basis in the absence of exposure level switching. [‡]95% CI refers to the pointwise 95% confidence interval $\hat{\mu} \pm 1.96\sqrt{\hat{\sigma}^2/5000}$, where $\hat{\mu}$ denotes the mean estimated risk difference and $\hat{\sigma}^2$ its empirical variance, i.e., the sample variance of the sample of 5000 estimates

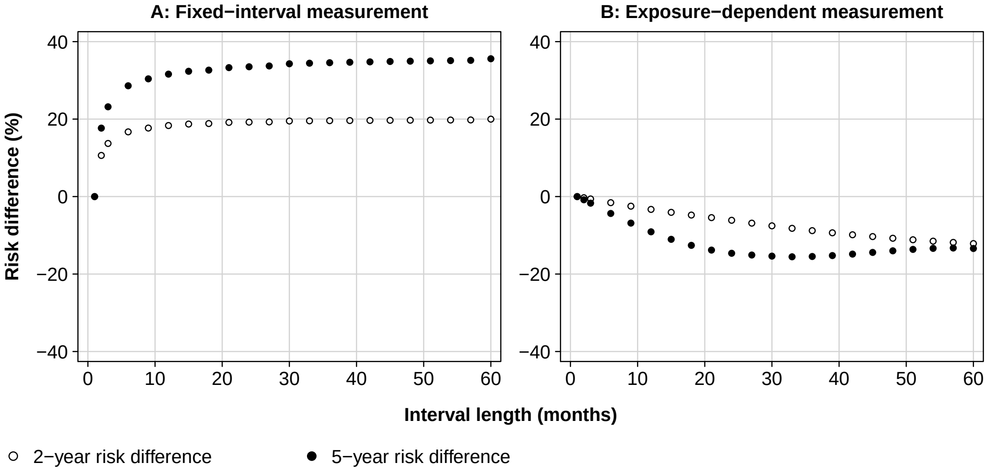


**Figure S7.2:** Mean estimated two- and five-year event risk differences across 5000 samples of size 150 000. Estimates derive from observational studies with varying covariate measurement designs. Panel A gives the estimates for fixed-interval measurement; panel B gives the estimates for covariate measurement with exposure switching and with fixed-length intervals in periods without switching.

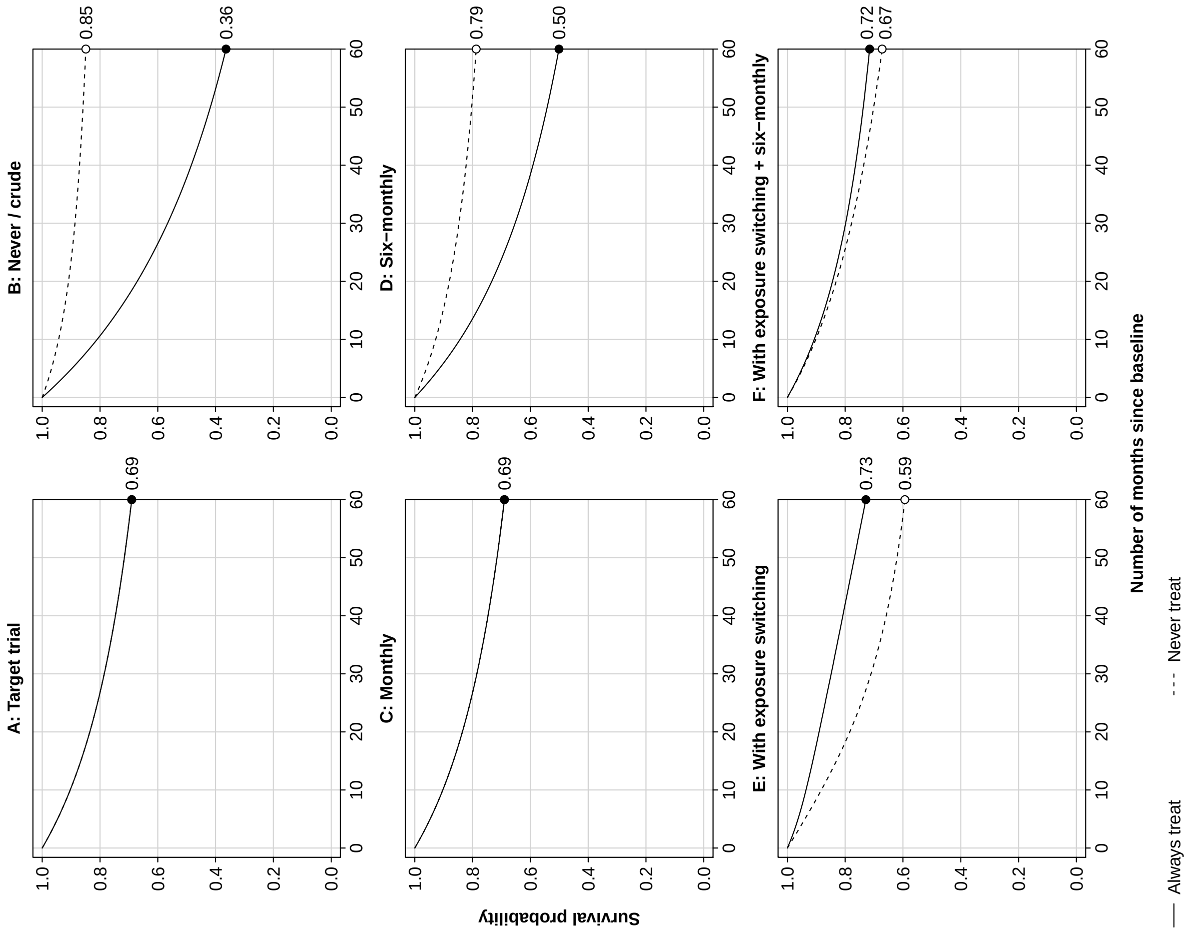


**Figure S7.3:** Mean estimated event-free survival probabilities across 5000 samples of size 10 000 based on target trial (panel A) and observational study (B through F) with varying covariate measurement designs: no covariate measurement (B), continuous to monthly covariate measurement (C), six-monthly covariate measurement (D), covariate measurement only with covariate level switching (E), and with exposure switching and six-monthly in periods without switching (F).

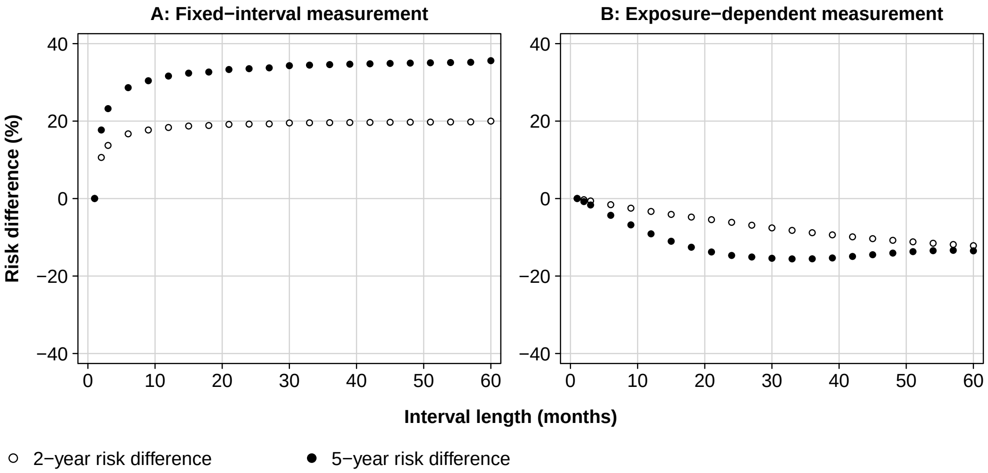


**Figure S7.4:** Mean estimated two- and five-year event risk differences across 5000 samples of size 10 000. Estimates derive from observational studies with varying covariate measurement designs. Panel A gives the estimates for fixed-interval measurement; panel B gives the estimates for covariate measurement with exposure switching and with fixed-length intervals in periods without switching.

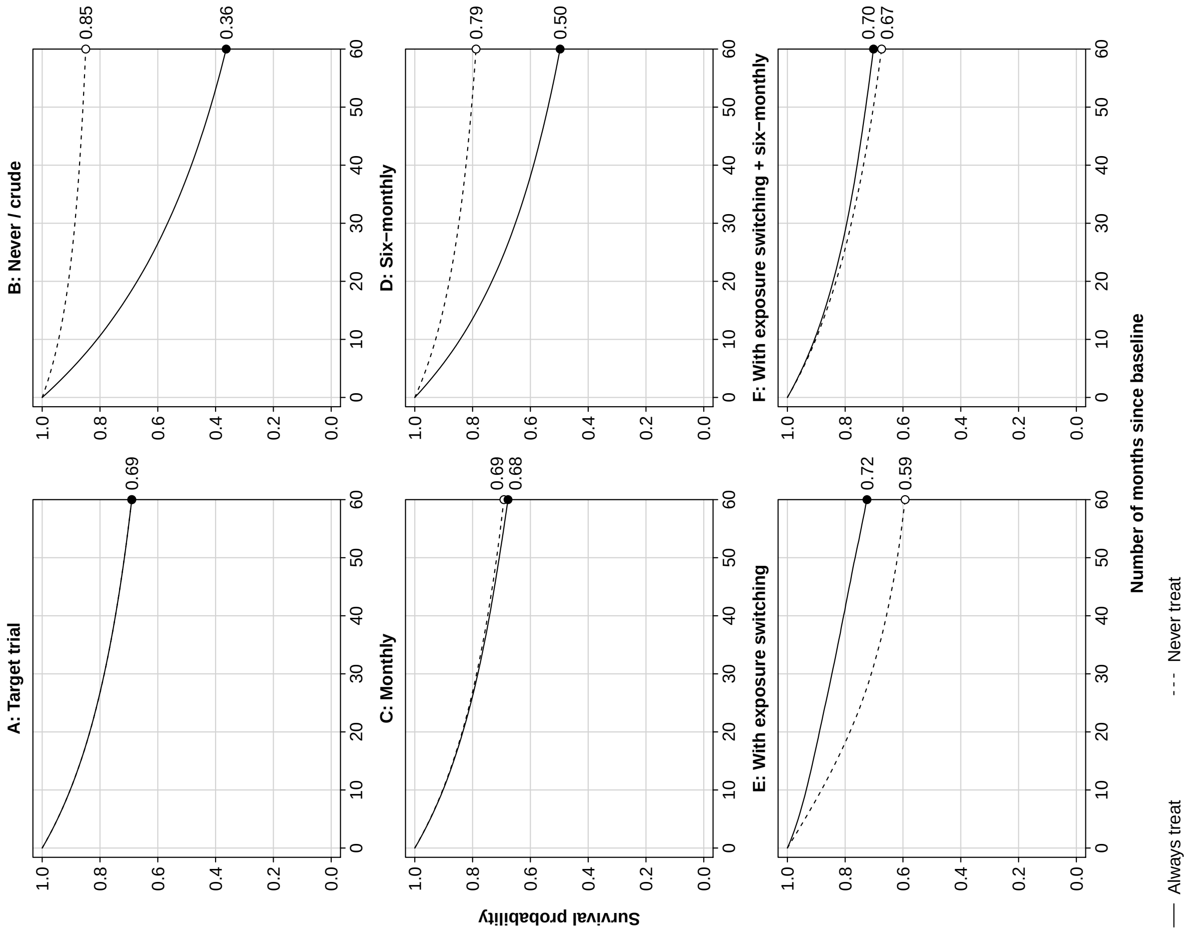


**Figure S7.5:** Mean estimated event-free survival probabilities across 5000 samples of size 1000 based on target trial (panel A) and observational study (B through F) with varying covariate measurement designs: no covariate measurement (B), continuous to monthly covariate measurement (C), six-monthly covariate measurement (D), covariate measurement only with covariate level switching (E), and with exposure switching and six-monthly in periods without switching (F).

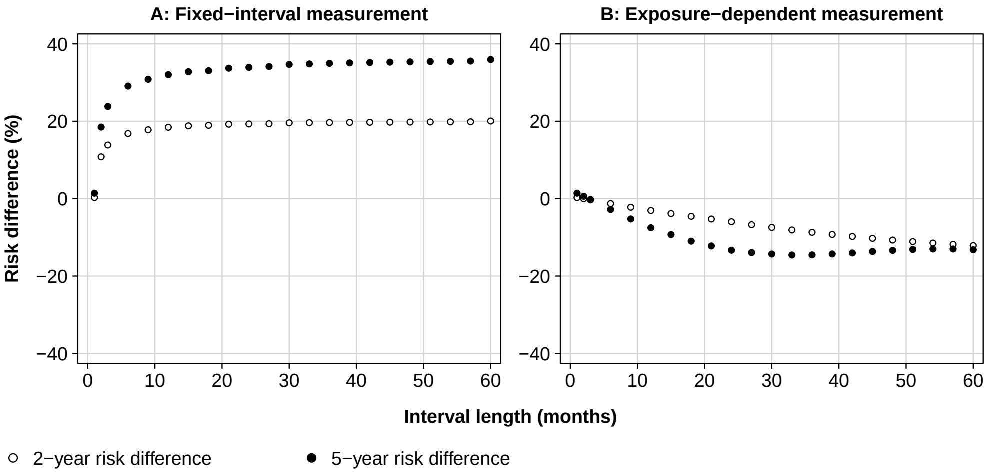


**Figure S7.6:** Mean estimated two- and five-year event risk differences across 5000 samples of size 1000. Estimates derive from observational studies with varying covariate measurement designs. Panel A gives the estimates for fixed-interval measurement; panel B gives the estimates for covariate measurement with exposure switching and with fixed-length intervals in periods without switching.

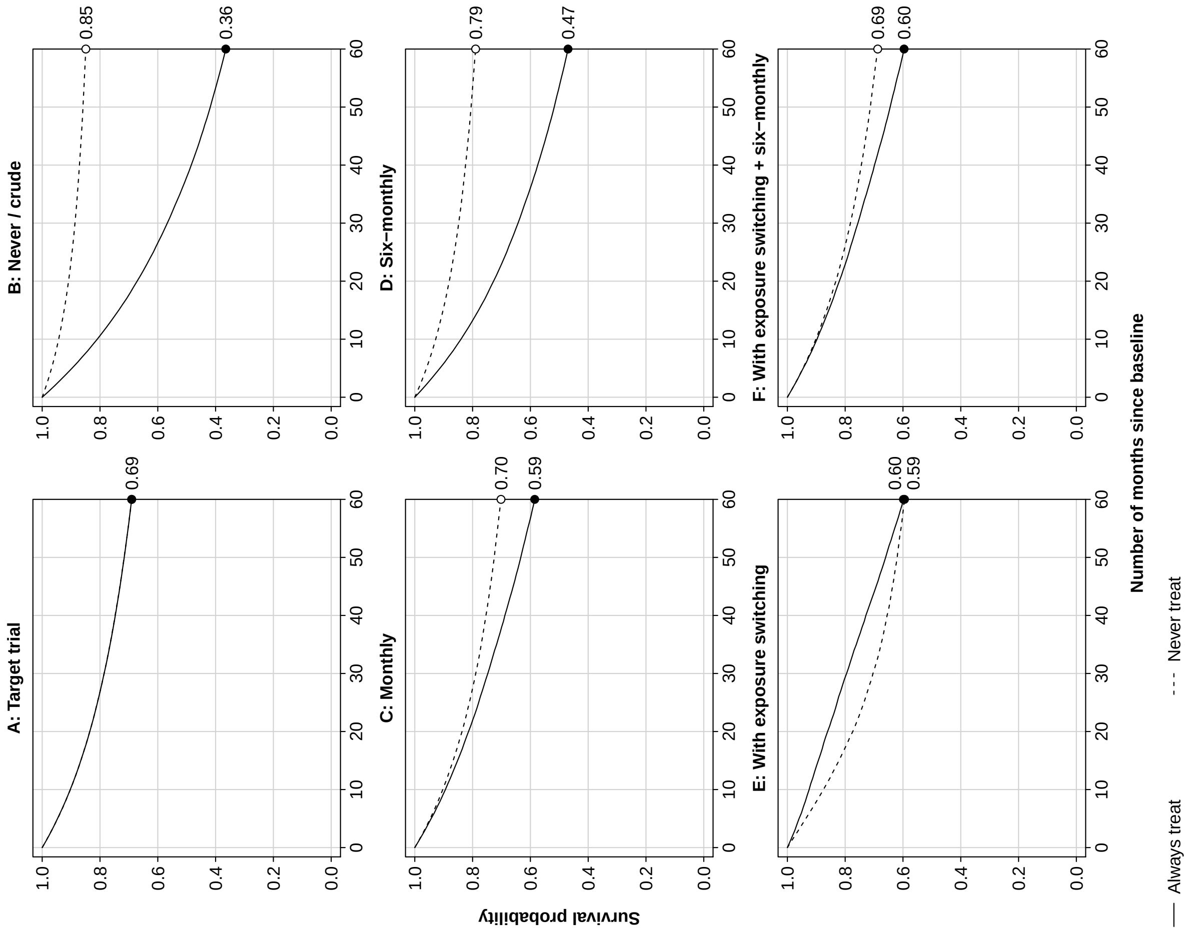


**Figure S7.7:** Mean estimated event-free survival probabilities across 5000 samples of size 100 based on target trial (panel A) and observational study (B through F) with varying covariate measurement designs: no covariate measurement (B), continuous to monthly covariate measurement (C), six-monthly covariate measurement (D), covariate measurement only with covariate level switching (E), and with exposure switching and six-monthly in periods without switching (F).

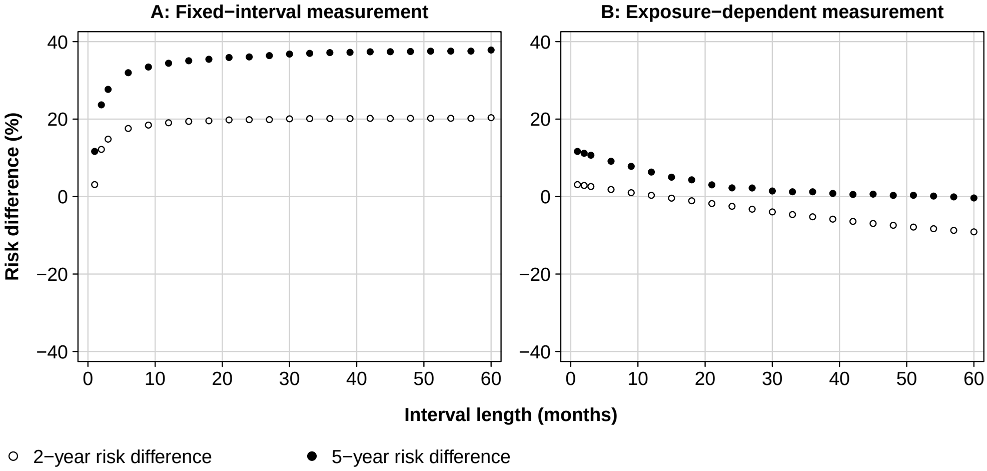


**Figure S7.8:** Mean estimated two- and five-year event risk differences across 5000 samples of size 100. Estimates derive from observational studies with varying covariate measurement designs. Panel A gives the estimates for fixed-interval measurement; panel B gives the estimates for covariate measurement with exposure switching and with fixed-length intervals in periods without switching.

Chapter 7

# Supplementary R code

```
# R code to supplement ’Bias of time-varying exposure effects due to
#   time-varying covariate measurement strategies’
# Compiled by Bas B.L. Penning de Vries (last updated: 7 Dec 2020)

# ==================================================================
# Preliminaries
# ==================================================================

# settings
K <- 60L # maximum number of months of follow-up
n <- 1.5e5L # sample size

# useful functions:
expit <- function(x) 1/(1+exp(-x))
locf <- function(x){
  # Last Observation Carried Forward
  isNA <- is.na(x)
  if(isNA[1L]) stop(’the first element is NA.’)
  y <- rep(x[!isNA],tabulate(cumsum(!isNA)))
  return(y)
}
qFirst <- function(x,last=FALSE){
  # Tests whether elements in x are the first occurrence of the
  #   corresponding values
  if(last) x <- rev(x)
  n <- length(x)
  w <- seq_len(n)
  o <- order(x)
  x <- x[o]
  y <- c(TRUE,x[-1L]!=x[-n])
  z <- y[match(w,o)]
  if(last) z <- rev(z)
  return(z)
}

# ==================================================================
# Data generating mechanism
# ==================================================================

drawSample <- function(n,trial=FALSE){
  sq <- seq(0,K-1e-6)
  A <- L <- matrix(nrow=n,ncol=length(sq))
  colnames(A) <- paste0("A",sq)
  colnames(L) <- paste0("L",sq)
  U <- runif(n)
  lagL <- lagA <- rep(0L,n)
  S <- rep(0,n)
  for(j in seq_along(sq)){
    Surv <- S>=(j-1L)
    Lk <- ifelse(Surv,runif(n)<expit(-.5+.25*(j==1L)+6*(U-.5)+
      .5*(lagA-.5)+1*(lagL-.5)),NA)
    g <- if(trial) {if(j==1L) rep(.5,n) else lagA} else
```

```
      expit(4*(j==1L)+10*(lagA-.5)+4*(Lk-.5))
    Ak <- ifelse(Surv,runif(n)<g,NA)
    L[,j] <- Lk
    A[,j] <- Ak
    eta <- 7*(U-.5)
    s <- suppressWarnings(rexp(n,rate=exp(-6+eta)))
    s[is.na(s)] <- 0
    S <- S+Surv*pmin(s,1L)
    Surv <- ifelse(Surv,s>1L,FALSE)
    lagL <- Lk; lagA <- Ak
  }
  status <- S<K
  S[!status] <- K
  return(data.frame(L*1L,A*1L,S=S,status=status))
}
coarsen <- function(data,design="I",after=6L){
  # assumes 'data' to be in long format
  out <- switch(design,
    I={ # exposure switch
      lagA <- ifelse(qFirst(data$unit),0L,c(0L,data$A[-nrow(data)]))
      M <- (data$start!=0L)&data$A==lagA&data$start>=0L
      data$L[M] <- NA
      return(data)
    },
    II={
    # set to missing if not multiple of 'after' months from baseline
      M <- data$start%%after!=0L&data$start>=0L
      data$L[M] <- NA
      return(data)
    },
    III={
    # set to missing if no switch of exposure AND not multiple of
    #   'after' months from since last exposure switch
      lagA <- ifelse(qFirst(data$unit),0L,c(0L,data$A[-nrow(data)]))
      M <- (data$start!=0L)&data$A==lagA&data$start>=0L
      cs <- cumsum(M)
      wh <- cumsum(!M)
      qf <- qFirst(wh)
      monthsSinceLastSwitch <- cs-cs[qf][match(wh,unique(wh))]
      M[!monthsSinceLastSwitch%%after] <- FALSE
      data$L[M] <- NA
      return(data)
    }
  )
}

# ==============================================================
# Data pre-processing functions
# ==============================================================

longFormat <- function(data){
  n <- nrow(data)
  w_L <- grep("^L-[:0-9:]+$",colnames(data))
  wL <- grep("^L[:0-9:]+$",colnames(data))
  wA <- grep("^A[:0-9:]+$",colnames(data))
  unit <- matrix(seq_len(n),nrow=n,ncol=length(wA),
    byrow=FALSE)[!is.na(data[,wA])]
  column <- matrix(seq_along(wA),nrow=n,ncol=length(wA),
    byrow=TRUE)[!is.na(data[,wA])]
  w <- cbind(unit,column)[order(unit),]
  out <- data.frame(unit=w[,1L],start=w[,2L]-1L)
  out$stop <- out$start+1L
  out$stop[qFirst(out$unit,TRUE)] <- data$S
  out$L <- data[,wL][w]
  out$A <- data[,wA][w]
  out$event <- FALSE
  out$event[qFirst(out$unit,TRUE)] <- data$status
  rownames(out) <- NULL
  return(out)
}
lagVariables <- function(data,m=1L){
  lagL <- matrix(nrow=nrow(data),ncol=m)
  colnames(lagL) <- paste0("lag",seq_len(m),"L")
  lagA <- matrix(nrow=nrow(data),ncol=m)
  colnames(lagA) <- paste0("lag",seq_len(m),"A")
  wh <- which(colnames(data)%in%c(colnames(lagL),colnames(lagA)))
  if(any(wh)) data <- data[,-wh,drop=FALSE]
  record <- data$start-min(data$start)+1L
  for(i in seq_len(m)){
    lagL[,i] <- ifelse(record>i,c(rep(0L,i),
      data$L[-(nrow(data)-0:(i-1))]),0L)
    lagA[,i] <- ifelse(record>i,c(rep(0L,i),
      data$A[-(nrow(data)-0:(i-1))]),0L)
  }
  data <- cbind(data,lagL,lagA)
  return(data)
}
LOCF <- function(data){
  data$L <- locf(data$L)
  return(data)
}
qAdhering <- function(data){
  switched <-
    ifelse(qFirst(data$unit),TRUE,c(TRUE,diff(data$A)!=0L))*1L
  cs <- cumsum(switched)
  mt <- with(data,match(unit,unique(unit)))
  return(cs==cs[qFirst(data$unit)][mt])
}

# ==================================================================
# Estimators
# ==================================================================

getPS <- function(data){
  fit <- glm(A~I(!start)+lag1A+L,data=data[data$start>=0L,,
    drop=FALSE],family=binomial)
  return(unname(predict(fit,newdata=data,type="response")))
}
estimateIPW <- function(data,ps){
  data$ps <- ps
  data <- data[data$start>=0L,]
  data <- data[qAdhering(data),]
  lp <- log(ifelse(data$A>0L,data$ps,1-data$ps))
  cs <- cumsum(lp)
  ql <- qFirst(data$unit,TRUE)
  data$W <- 1/exp(cs-c(0,cs[ql][-sum(ql)])[match(data$unit,
    unique(data$unit))])
  EW0 <- with(data[data$A==0L,],tapply(W,start,mean))
  EW1 <- with(data[data$A==1L,],tapply(W,start,mean))
  mt0 <- with(data,match(start,unique(start)))
  mt1 <- with(data,match(start,unique(start)))
  data$sW <- data$W/ifelse(data$A>0L,EW1[mt1],EW0[mt0])
  fit <- with(data,survival::survfit(survival::Surv(start,stop,
    event)~A,weights=sW,timefix=FALSE))
  smmry <- survival:::summary.survfit(fit,times=0:K)
  est <- split(smmry$surv,smmry$strata)
  names(est) <- c("surv0","surv1")
  est$surv0 <- c(est$surv0,rep(rev(est$surv0)[1L],
    K+1L-length(est$surv0)))
  est$surv1 <- c(est$surv1,rep(rev(est$surv1)[1L],
    K+1L-length(est$surv1)))
  return(est)
}
crudePP <- function(data){
  data <- data[data$start>=0L,]
  data <- data[qAdhering(data),,drop=FALSE]
  ql <- qFirst(data$unit,TRUE)
  time <- data$stop[ql]
  status <- data$event[ql]
  group <- data$A[ql]
  fit <- survival::survfit(survival::Surv(time,status)~group)
  smmry <- survival:::summary.survfit(fit,times=0:K)
  est <- split(smmry$surv,smmry$strata)
  names(est) <- c("surv0","surv1")
  est$surv0 <- c(est$surv0,rep(rev(est$surv0)[1L],
    K+1L-length(est$surv0)))
  est$surv1 <- c(est$surv1,rep(rev(est$surv1)[1L],
    K+1L-length(est$surv1)))
  return(est)
}

# ============================================================
# Data generation & estimation
# ============================================================

trial <- longFormat(drawSample(n,trial=TRUE))
wide <- drawSample(n)
long <- lagVariables(longFormat(wide))
longI <- lagVariables(LOCF(coarsen(long,"I")))
longII <- lagVariables(LOCF(coarsen(long,"II")))
longIII <- lagVariables(LOCF(coarsen(long,"III")))

estA <- crudePP(trial)
estB <- crudePP(long)
estC <- estimateIPW(long,getPS(long))
```

```
estD <- estimateIPW(longII,getPS(longII))
estE <- estimateIPW(longI,getPS(longI))
estF <- estimateIPW(longIII,getPS(longIII))

sq <- c(1L,2L,seq(3L,60L,3))
estVarD <- estVarF <- list()
for(i in seq_along(sq)){
  cat("\r",i,"/",length(sq),sep=""); flush.console()
  longIIi <- lagVariables(LOCF(coarsen(long,"II",after=sq[i])))
  estVarD[[i]] <- estimateIPW(longIIi,getPS(longIIi))
  longIIIi <- lagVariables(LOCF(coarsen(long,"III",after=sq[i])))
  estVarF[[i]] <- estimateIPW(longIIIi,getPS(longIIIi))
}
names(estVarD) <- names(estVarF) <- sq
```

# 8

# A weighting method for simultaneous adjustment for confounding and joint exposure-outcome misclassifications

Bas B. L. Penning de Vries
Maarten van Smeden
Rolf H. H. Groenwold

*Statistical Methods in Medical Research* 2021; 30(2): 473–487

## Abstract

Joint misclassification of exposure and outcome variables can lead to considerable bias in epidemiological studies of causal exposure-outcome effects. In this paper, we present a new maximum likelihood based estimator for marginal causal effects that simultaneously adjusts for confounding and several forms of joint misclassification of the exposure and outcome variables. The proposed method relies on validation data for the construction of weights that account for both sources of bias. The weighting estimator, which is an extension of the outcome misclassification weighting estimator proposed by Gravel and Platt (Statistics in Medicine, 2018), is applied to reinfarction data. Simulation studies were carried out to study its finite sample properties and compare it with methods that do not account for confounding or misclassification. The new estimator showed favourable large sample properties in the simulations. Further research is needed to study the sensitivity of the proposed method and that of alternatives to violations of their assumptions. The implementation of the estimator is facilitated by a new R function (`ipwm`) in an existing R package (`mecor`).

Chapter 8

## 8.1 Introduction

In epidemiological research on causal associations between a particular exposure and a certain outcome, erroneous information on either or both of these variables poses a serious methodological obstacle in making valid inferences. In particular, joint misclassification of exposure and outcome can lead to considerable bias of standard causal effect estimators, with direction and magnitude depending on various factors, including the misclassification mechanism and the direction and magnitude of the true effect (Kristensen, 1992; Brenner et al., 1993; Vogel et al., 2005; Jurek et al., 2008; VanderWeele and Hernán, 2012; Brooks et al., 2018).

Exposure and outcome misclassification is typically categorised according to two separate properties: whether or not the misclassification is differential and whether or not it is dependent relative to some covariate vector $L$ containing patient characteristics (Kristensen, 1992; VanderWeele and Hernán, 2012). Joint misclassification of exposure and outcome is said to be *nondifferential* if (1) the sensitivity and specificity of exposure classification are constant across all categories of the (true) outcome given $L$ and (2) the sensitivity and specificity of outcome classification are constant across all categories of the (true) exposure given $L$; otherwise it is *differential*. Misclassification is said to be *independent* if the joint probability of any exposure and outcome classification given any true exposure and outcome categories and $L$ can be factored into the product of the corresponding probabilities for exposure and outcome separately; otherwise, it is *dependent*. In Dawid's notation (1979), that is, if true exposure level $A$ and true outcome $Y$ are (potentially mis)classified as $B$ and $Z$, respectively, misclassification is nondifferential if and only if $B \perp\!\!\!\perp Y | A, L$ and $Z \perp\!\!\!\perp A | Y, L$ and independent if and only if $Z \perp\!\!\!\perp B | Y, A, L$.

Epidemiological research hampered by joint misclassification of some type is likely voluminous (Brooks et al., 2018). Examples of studies affected by exposure and outcome misclassification can be found, for example, in the literature on the causal effects of drug use, which is largely based on routinely collected data, where exposures are typically operationalised on the basis of prescription records and where outcomes are often self-reported (Marcum et al., 2013; Culver et al., 2012; Leong et al., 2013; Ni et al., 2017). In applied epidemiological research, misclassification or some of its potential consequences are often ignored (Jurek et al., 2006; Brakenhoff et al., 2018). The assertion often made in the discussion of study results that observed measures of association are biased toward the null under nondifferentiality, for example, is not generally true unless additional conditions are presupposed (Brenner et al., 1993; Brooks et al., 2018).

Methods to adjust for misclassification rely on additional information that can be used to estimate or correct for bias. One potential source of information is validation data obtained through supposedly infallible measurement. Recently, Gravel and Platt (2018) proposed an inverse probability weighting (IPW) method to simultaneously address confounding and outcome misclassification by means of internal validation data. Other methods likewise suppose that either the exposure or the outcome is subject to misclassification (Babanezhad et al., 2010; Braun et al., 2017; Gravel and Platt, 2018; Shu and Yi, 2019). In what follows, we propose an extension of Gravel and Platt's method to allow for confounding adjustment and joint exposure and outcome misclassification. This flexible estimator allows for the misclassifications to be dependent, differential or both. In Section 8.2, inverse probability weights for confounding and joint misclassification are introduced through a hypothetical study based on the illustrative example of Gravel and Platt. Section 8.3 details methods for estimation of the various components of the proposed weights based on validation data. In Section 8.4, we describe a series of Monte Carlo simulations that were used to study properties of the proposed method in finite samples. We conclude with a summary and discussion of our findings in context of the existing literature.

## 8.2 Data distribution for illustration and development of weighting method

We first consider the data and setting described by Gravel and Platt and suppose that Table 8.1 represents a simple random (i.i.d.) sample from (or that its cell counts are proportional to the respective densities in) the population of interest. This illustration is based on a cohort study on the association between post-myocardial infarction statin use ($A$) and the 1-year risk of reinfarction ($Y$). In what follows, we will refer to this example as the 'reinfarction example'.

Throughout we take the counterfactual framework for causal inference, formal accounts of which are given for example by Neyman, Rubin, Holland, and Pearl (Neyman et al., 1935; Rubin, 1974; Holland, 1986, 1988; Pearl, 2009). The interest, we suppose, lies in estimating $g(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$ for some function $g$, where $Y(0)$ and $Y(1)$ denote the counterfactual outcomes for hypothetical interventions setting $A$ to 0 and 1, respectively. Common choices of $g$ define $g(p_0, p_1) = p_1 - p_0$ (risk difference), $g(p_0, p_1) = p_1/p_0$ (risk ratio) or $g(p_0, p_1) = [p_1/(1-p_1)]/[p_0/(1-p_0)]$ (odds ratio). For our numerical example and simulation studies, we concentrate on the causal marginal odds ratio (OR) in particular, with

$$\text{OR} = g(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)]) = \frac{\mathbb{E}[Y(1)]/(1 - \mathbb{E}[Y(1)])}{\mathbb{E}[Y(0)]/(1 - \mathbb{E}[Y(0)])}, \tag{8.1}$$

but the results naturally extend to other effect measures.

### *8.2.1 No misclassification*

Under conditional exchangeability given $L$ (i.e., $(Y(0), Y(1)) \perp\!\!\!\perp A|L$), consistency ($Y(a) = Y$ if $A = a$) and positivity ($\Pr(A = a|L = l) > 0$ for $a = 0, 1$ and all $l$ in the support of $L$), the mean counterfactuals $E[Y(0)]$ and $E[Y(1)]$ can be expressed in terms of 'observables' (meaning, here, variables that would be observed in the absence of measurement error) as follows:

$$\mathbb{E}[Y(0)] = \mathbb{E}[WY|A = 0] \text{ and } \mathbb{E}[Y(1)] = \mathbb{E}[WY|A = 1],$$

where $W$ denotes the inverse probability of the allocated exposure level $A$ given $L$ multiplied by the prevalence of the allocated exposure level $A$ (i.e., $W = \Pr(A)/\Pr(A|L)$; Supplementary Appendix S8.1). We therefore have

$$g(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)]) = g(\mathbb{E}[WY|A = 0], \mathbb{E}[WY|A = 1]). \tag{8.2}$$

Replacing components of the right-hand side of (8.2) with sample analogues, we obtain the following estimator for the setting where $L$ is binary:

$$\begin{aligned}\widehat{\mathrm{OR}} &:= g(\widehat{\mathbb{E}}[\widehat{W}Y|A = 0], \widehat{\mathbb{E}}[\widehat{W}Y|A = 1])\\ &= \frac{\widehat{\mathbb{E}}[\widehat{W}Y|A = 1]/(1 - \widehat{\mathbb{E}}[\widehat{W}Y|A = 1])}{\widehat{\mathbb{E}}[\widehat{W}Y|A = 0]/(1 - \widehat{\mathbb{E}}[\widehat{W}Y|A = 0])}\\ &= \frac{(\widehat{W}_{10}n_{110} + \widehat{W}_{11}n_{111})/(n_{110} + n_{111} + n_{010} + n_{011} - \widehat{W}_{10}n_{110} - \widehat{W}_{11}n_{111})}{(\widehat{W}_{00}n_{100} + \widehat{W}_{01}n_{101})/(n_{100} + n_{101} + n_{000} + n_{001} - \widehat{W}_{00}n_{100} - \widehat{W}_{01}n_{101})},\end{aligned} \tag{8.3}$$

where $n_{yal}$ denotes the number of subjects with $Y = y$, $A = a$, $L = l$ and where $\widehat{W}_{al}$ is the product of the proportion of subjects in the sample with $A = a$ and

**Table 8.1:** Cross-classification of the reinfarction data for 33,007 individuals as given by Gravel and Platt (2018).

| | $L = 0$ | | $L = 1$ | |
|---|---|---|---|---|
| | $A = 0$ | $A = 1$ | $A = 0$ | $A = 1$ |
| $Y = 0$ | 11602 | 13116 | 1302 | 5363 |
| $Y = 1$ | 890 | 589 | 49 | 96 |

the inverse of the proportion of subjects with $A = a$ among those with $L = l$. For the data in Table 8.1, we obtain $\widehat{\mathrm{OR}} \approx 0.573$. The corresponding crude odds ratio (i.e., with $\widehat{W} = 1$) is 0.509.

### *8.2.2 Joint misclassification*

Suppose that rather than observing $Y$ and $A$ we observe $Z$ and $B$, the misclassified versions of $Y$ and $A$, respectively. The relation between $Z$ and $B$ on the one hand and $Y$, $A$ and $L$ on the other can be expressed as follows:

$$\begin{aligned}&\Pr(Z = z, B = b|Y = y, A = a, L = l)\\ &\quad = (\pi_{byal})^z(1-\pi_{byal})^{1-z}(\lambda_{yal})^b(1-\lambda_{yal})^{1-b}\end{aligned}$$

for $z, b \in \{0, 1\}$ and all possible realisations $y, a, l$ of $Y, A, L$, and where $\pi_{byal} = \Pr(Z = 1|B = b, Y = y, A = a, L = l)$ and $\lambda_{yal} = \Pr(B = 1|Y = y, A = a, L = l)$.

To simulate (dependent differential) misclassification in the reinfarction dataset, we use the true positive and false positive rates given in Table 8.2. The expected cell counts for these rates are given in Table 8.3.

We redefine the weights in (8.2) as a function of $B$ and $L$ (as per Supplementary Appendix S8.1) such that

$$W = \frac{p(B)\varepsilon_{BL}}{\sum_y \sum_a \pi_{ByaL}(\lambda_{yaL})^B(1-\lambda_{yaL})^{1-B}(\varepsilon_{aL})^y(1-\varepsilon_{aL})^{1-y}(\delta_L)^a(1-\delta_L)^{1-a}}, \tag{8.4}$$

where $p(B)$ is the prevalence of level $B$ of the potentially misclassified version of the exposure variable and where $\varepsilon_{al} = \Pr(Y = 1|A = a, L = l)$ and $\delta_l =$

**Table 8.2:** True and false positive rates for reinfarction example. For $b, y, a, l \in \{0, 1\}$, $\lambda_{yal} = \Pr(B = 1|Y = y, A = a, L = l)$ and $\pi_{byal} = \Pr(Z = 1|B = b, Y = y, A = a, L = l)$.

| | | |
|---|---|---|
| $\pi_{0000} = 0.050$ | $\pi_{0001} = 0.020$ | $\lambda_{000} = 0.010$ |
| $\pi_{1000} = 0.060$ | $\pi_{1001} = 0.108$ | $\lambda_{100} = 0.181$ |
| $\pi_{0100} = 0.930$ | $\pi_{0101} = 0.806$ | $\lambda_{010} = 0.880$ |
| $\pi_{1100} = 0.938$ | $\pi_{1101} = 0.692$ | $\lambda_{110} = 0.910$ |
| $\pi_{0010} = 0.030$ | $\pi_{0011} = 0.109$ | $\lambda_{001} = 0.100$ |
| $\pi_{1010} = 0.060$ | $\pi_{1011} = 0.050$ | $\lambda_{101} = 0.265$ |
| $\pi_{0110} = 0.906$ | $\pi_{0111} = 0.765$ | $\lambda_{011} = 0.930$ |
| $\pi_{1110} = 0.950$ | $\pi_{1111} = 0.861$ | $\lambda_{111} = 0.823$ |

$\Pr(A = 1|L = l)$ for all possible realisations $a$ and $l$ of $A$ and $L$, respectively. In Supplementary Appendix S8.1, it is shown that

$$\mathbb{E}[Y(0)] = \mathbb{E}[WZ|B = 0] \quad \text{and} \quad \mathbb{E}[Y(1)] = \mathbb{E}[WZ|B = 1], \tag{8.5}$$

which suggests the plug-in estimator

$$\begin{aligned} \widehat{\text{OR}} &:= g(\widehat{\mathbb{E}}[\widehat{W}Z|B = 0], \widehat{\mathbb{E}}[\widehat{W}Z|B = 1]) \\ &= \frac{\widehat{\mathbb{E}}[\widehat{W}Z|B = 1]/(1 - \widehat{\mathbb{E}}[\widehat{W}Z|B = 1])}{\widehat{\mathbb{E}}[\widehat{W}Z|B = 0]/(1 - \widehat{\mathbb{E}}[\widehat{W}Z|B = 0])}, \end{aligned} \tag{8.6}$$

where $\widehat{\mathbb{E}}$ denotes the sample mean operator and $\widehat{W}$ the sample analogue (i.e., consistent estimator) of $W$ in (8.4). For other effect measures (i.e., other choices of $g$), the same plug-in strategy can be implemented.

In the absence of exposure misclassification, (8.4) reduces to

$$W = \left( \frac{(\delta_L)^A (1 - \delta_L)^{1-A}}{p(A)} \left[ \pi_{A0AL} \frac{1 - \varepsilon_{AL}}{\varepsilon_{AL}} + \pi_{A1AL} \right] \right)^{-1}. \tag{8.7}$$

The first term within the round brackets corrects for confounding and represents the propensity of the received exposure level $A$ divided by prevalence of exposure level $A$. The term within square brackets is a factor that corrects for misclassification in the outcome variable. This correction factor is similar to

**Table 8.3:** Expected cell counts (rounded to integers) for reinfarction example after misclassification was introduced. Because of rounding, the sum of all cell entries is 33,006 rather than 33,007, the size of the reinfarction dataset.

| | $L = 0$ | | $L = 1$ | |
|---|---|---|---|---|
| | $A = 0$ | $A = 1$ | $A = 0$ | $A = 1$ |
| $Y = 0,\ A = 0,\ L = 0$ | 10912 | 109 | 574 | 7 |
| $Y = 1,\ A = 0,\ L = 0$ | 51 | 10 | 678 | 151 |
| $Y = 0,\ A = 1,\ L = 0$ | 1527 | 10850 | 47 | 693 |
| $Y = 1,\ A = 1,\ L = 0$ | 5 | 27 | 48 | 509 |
| $Y = 0,\ A = 0,\ L = 1$ | 1148 | 116 | 23 | 14 |
| $Y = 1,\ A = 0,\ L = 1$ | 7 | 4 | 29 | 9 |
| $Y = 0,\ A = 1,\ L = 1$ | 334 | 4738 | 41 | 249 |
| $Y = 1,\ A = 1,\ L = 1$ | 4 | 11 | 13 | 68 |

that proposed by Gravel and Platt (2018). The only difference is that where in (8.7) it does not depend on the fallible measurement $Z$ of $Y$, Gravel and Platt define different weights for subjects with $Z = 0$. Note, however, that the choice of weights for subjects with $Z = 0$ does not affect the population quantity in (8.5) or the estimator defined by (8.6), because the weights only appear in products with $Z$, which equal zero if $Z = 0$.

As for the reinfarction example, the odds ratio estimate for the exposure-outcome effect based on inverse probability weighting that assumes absence of exposure or outcome misclassification is 1.120, while the corresponding misclassification naive crude odds ratio is 1.031. Estimation of the population weights $W$ from observables using validation data is discussed in the next section. As shown below, weighting using the proposed weights that account for confounding and outcome and exposure misclassification results in an odds ratio of $\mathrm{OR} = \widehat{\mathrm{OR}} \approx 0.573$. Inference based on (8.7) rather than (8.4), i.e., using Gravel and Platt's method and ignoring misclassification in the exposure but correcting for outcome misclassification, yields an odds ratio estimate of 0.934.

### *8.2.3 Parameterisation based on positive and negative predictive values*

In the foregoing discussion, the proposed weights were expressed in terms of sensitivity and specificity parameters. The sensitivity and specificity of $Z$ with respect to $Y$, given $(B, A, L)$, are $\pi_{B1AL}$ and $1 - \pi_{B0AL}$, respectively. Similarly, $\lambda_{Y1L}$ and $1 - \lambda_{Y0L}$ reflect the sensitivity and specificity, respectively, with respect to $A$, conditional on $Y$ and $L$.

As discussed below, it may be more convenient to choose a parameterisation that is based on (positive and negative) predictive values. Define $\delta^*_l = \Pr(B = 1|L = l)$, $\varepsilon^*_{bl} = \Pr(Z = 1|B = b, L = l)$, $\lambda^*_{zbl} = \Pr(A = 1|Z = z, B = b, L = l)$ and $\pi^*_{azbl} = \Pr(Y = 1|A = a, Z = z, B = b, L = l)$. The weights in (8.4) can be rewritten as

$$W = \frac{\sum_y \sum_a \pi^*_{ByaL}(\lambda^*_{yaL})^B(1-\lambda^*_{yaL})^{1-B}(\varepsilon^*_{aL})^y(1-\varepsilon^*_{aL})^{1-y}(\delta^*_L)^a(1-\delta^*_L)^{1-a}}{\sum_y \sum_a (\lambda^*_{yaL})^B(1-\lambda^*_{yaL})^{1-B}(\varepsilon^*_{aL})^y(1-\varepsilon^*_{aL})^{1-y}(\delta^*_L)^a(1-\delta^*_L)^{1-a}} \times \frac{p(B)}{\varepsilon^*_{BL}(\delta^*_L)^B(1-\delta^*_L)^{1-B}}. \tag{8.8}$$

In the absence of exposure misclassification, these weights simplify to

$$W = \frac{p(A)}{(\delta_L)^A(1-\delta_L)^{1-A}} \frac{\varepsilon_{AL}}{\varepsilon^*_{AL}}.$$

## 8.3 Estimation of weights based on validation data

Estimation of the proposed weights can be done using a number of approaches and we will here consider a maximum likelihood approach that assumes the availability of internal validation data, i.e., that some study participants have their observed exposure or outcome measured by an 'infallible' or 'gold standard' (100% accurate) classifier, and that all participants have the misclassified exposure and outcome variables measured.

### *8.3.1 Validation subset inclusion mechanism*

Let $R_Y$ be the indicator variable that takes the value of 1 if the outcome is observed (i.e., measured by an infallible classifier) and 0 otherwise. Similarly, define $R_A$ to be the indicator variable that takes the value of 1 if the exposure variable is observed and 0 otherwise. $R_Y$ and $R_A$ reflect which subjects have validation data available on $Y$ and $A$, respectively. The subset of subjects with validation data on $Y$ need not fully overlap with the subset with validation data on $A$.

The validation subsets can be approached from the missing data framework of Rubin (1976) Provided that $Z, B, L$ are free of missing values, Rubin's missing at random (MAR) condition is met if the vector $(R_Y, R_A)$ is conditionally independent of $(Y, A)$ given $(Z, B, L)$.

### *8.3.2 Full likelihood approach based on parameterisation in terms of sensitivities and specificities*

Simultaneous estimation of the whole vector of $\delta$, $\varepsilon$, $\lambda$ and $\pi$ parameters can be done via maximum likelihood estimation as follows. Assuming i.i.d. observations $(Z_i, B_i, Y_i, A_i, L_i)$ and ignorable missingness in the sense of Rubin (1976) (MAR and distinctness), for valid likelihood-based inference it is appropriate to maximise the following log-likelihood over the parameter space of $\theta$, the vector of $\delta$, $\varepsilon$, $\lambda$ and $\pi$ parameters:

$$\begin{aligned}\ell(\theta) = &\sum_{i:R_{Yi}=R_{Ai}=1} \log f(\theta; Z_i, B_i, Y_i, A_i, L_i) \\ &+ \sum_{i:R_{Yi}=1\wedge R_{Ai}=0} \log \sum_{A_i} f(\theta; Z_i, B_i, Y_i, A_i, L_i) \\ &+ \sum_{i:R_{Yi}=0\wedge R_{Ai}=1} \log \sum_{Y_i} f(\theta; Z_i, B_i, Y_i, A_i, L_i)\end{aligned}$$

$$+ \sum_{i:R_{Yi}=R_{Ai}=0} \log \sum_{Y_i} \sum_{A_i} f(\theta; Z_i, B_i, Y_i, A_i, L_i),$$

where

$$\begin{aligned} & f(\theta; Z_i, B_i, Y_i, A_i, L_i) \\ & \quad = (\pi_{B_iY_iA_iL_i})^{Z_i}(1-\pi_{B_iY_iA_iL_i})^{1-Z_i}(\lambda_{Y_iA_iL_i})^{B_i}(1-\lambda_{Y_iA_iL_i})^{1-B_i} \\ & \qquad \times (\varepsilon_{A_iL_i})^{Y_i}(1-\varepsilon_{A_iL_i})^{1-Y_i}(\delta_{L_i})^{A_i}(1-\delta_{L_i})^{1-A_i}. \end{aligned}$$

Evaluating this log-likelihood involves marginalising over unobserved quantities in the last three terms of $\ell(\theta)$. The log-likelihood equations may become considerably more tractable if we choose a parameterisation of the likelihood that is based on predictive values rather than sensitivities and specificities.

### *8.3.3 Full likelihood approach based on parameterisation in terms of predictive values*

Inference may alternatively be based on a log-likelihood that is parameterised in terms of the vector $\theta^*$ of the $\delta^*$, $\varepsilon^*$, $\lambda^*$ and $\pi^*$ parameters, i.e.,

$$\begin{aligned} \ell^*(\theta^*) = & \sum_{i:R_{Yi}=R_{Ai}=1} \log h(\theta^*; Z_i, B_i, Y_i, A_i, L_i) \\ & + \sum_{i:R_{Yi}=1 \wedge R_{Ai}=0} \log \sum_{A_i} h(\theta^*; Z_i, B_i, Y_i, A_i, L_i) \\ & + \sum_{i:R_{Yi}=0 \wedge R_{Ai}=1} \log \sum_{Y_i} h(\theta^*; Z_i, B_i, Y_i, A_i, L_i) \\ & + \sum_{i:R_{Yi}=R_{Ai}=0} \log \sum_{Y_i} \sum_{A_i} h(\theta^*; Z_i, B_i, Y_i, A_i, L_i), \end{aligned}$$

where

$$\begin{aligned} & h(\theta^*; Z_i, B_i, Y_i, A_i, L_i) \\ & \quad = (\pi^*_{A_iZ_iB_iL_i})^{Y_i}(1-\pi^*_{A_iZ_iB_iL_i})^{1-Y_i}(\lambda^*_{Z_iB_iL_i})^{A_i}(1-\lambda^*_{Z_iB_iL_i})^{1-A_i} \\ & \qquad \times (\varepsilon^*_{B_iL_i})^{Z_i}(1-\varepsilon^*_{B_iL_i})^{1-Z_i}(\delta^*_{L_i})^{B_i}(1-\delta^*_{L_i})^{1-B_i}. \end{aligned}$$

If validation data is available on $Y$ if and only if it is available on $A$, the complete data log-likelihood ignoring the missing data mechanism can be conveniently expressed as follows:

$$\ell^*(\theta^*) = \ell^*_1(\theta^*) + \ell^*_2(\theta^*) + \ell^*_3(\theta^*) + \ell^*_4(\theta^*), \tag{8.9}$$

with $\theta^*$ denoting the vector of $\delta^*$, $\varepsilon^*$, $\lambda^*$ and $\pi^*$ parameters and where

$$\begin{aligned}
\ell_1^*(\theta^*) &= \sum_{i:R_{Yi}=R_{Ai}=1} Y_i \log(\pi^*_{A_iZ_iB_iL_i}) + (1-Y_i)\log(1-\pi^*_{A_iZ_iB_iL_i}),\\
\ell_2^*(\theta^*) &= \sum_{i:R_{Yi}=R_{Ai}=1} A_i \log(\lambda^*_{Z_iB_iL_i}) + (1-A_i)\log(1-\lambda^*_{Z_iB_iL_i}),\\
\ell_3^*(\theta^*) &= \sum_i Z_i \log(\varepsilon^*_{B_iL_i}) + (1-Z_i)\log(1-\varepsilon^*_{B_iL_i}),\\
\ell_4^*(\theta^*) &= \sum_i B_i \log(\delta^*_{L_i}) + (1-B_i)\log(1-\delta^*_{L_i}).
\end{aligned}$$

Now, assuming distinct parameter spaces for the vectors of $\pi^*$, $\lambda^*$, $\varepsilon^*$, and $\delta^*$ parameters, the parameter values that maximise $\ell^*(\theta^*)$ can be found by separately maximising $\ell_1^*(\theta^*)$ and $\ell_2^*(\theta^*)$ in the validation subset with respect to the $\pi^*$ and $\lambda^*$ parameters, respectively, and $\ell_3^*(\theta^*)$ and $\ell_4^*(\theta^*)$ in the entire dataset with respect to $\varepsilon^*$ and $\delta^*$. Following Gravel and Platt (2018) and Tang et al. (2013), the sum of the first and last two terms are therefore suitably labelled the internal validation and main study log-likelihood, respectively. With this parameterisation, finding the maximum likelihood estimates is readily achieved by taking advantage of standard statistical software.

### *8.3.4 Equivalence of likelihood approaches based on different parameterisations*

Without restrictions imposed on

$$\begin{aligned}
\theta_l &:= (\pi_{000l}, \pi_{100l}, \pi_{010l}, \pi_{110l}, \pi_{001l}, \pi_{101l}, \pi_{011l}, \pi_{111l}, \lambda_{00l}, \lambda_{10l}, \lambda_{01l}, \lambda_{11l},\\
&\qquad \varepsilon_{0l}, \varepsilon_{1l}, \delta_l) \quad \text{and}\\
\theta_l^* &:= (\pi^*_{000l}, \pi^*_{100l}, \pi^*_{010l}, \pi^*_{110l}, \pi^*_{001l}, \pi^*_{101l}, \pi^*_{011l}, \pi^*_{111l}, \lambda^*_{00l}, \lambda^*_{10l}, \lambda^*_{01l}, \lambda^*_{11l},\\
&\qquad \varepsilon^*_{0l}, \varepsilon^*_{1l}, \delta^*_l),
\end{aligned}$$

other than that $\theta_l, \theta_l^* \in (0,1)^{15}$, it can be shown that the maximum likelihood estimator based on the internal validation design is invariant to its parameterisation (sensitivities/specificities versus positive and negative predictive values). This is because there exists a function mapping every $\theta_l \in (0,1)^{15}$ to a unique $\theta_l^* \in (0,1)^{15}$ and vice versa. Maximising $\ell(\theta)$ with respect to $\theta$ is then equivalent to maximising $\ell(\sigma(\theta^*))$ $(= \ell^*(\theta^*))$ with respect to $\theta^*$ for some bijection $\sigma$ such that $\theta = \sigma(\theta^*)$; that is,

$$\arg\max_{\theta}\ \ell(\theta) = \sigma\left(\arg\max_{\theta^*}\ \ell(\sigma(\theta^*))\right).$$

If more restrictions are imposed on $\theta$ or $\theta^*$, e.g., if we assume non-saturated logistic models for the components of $\theta$ and $\theta^*$, this equivalence no longer holds and the resulting weight estimates may differ depending on the parameterisation.

### *8.3.5 Application*

For the re-infarction data example, we assume validation data are available according to a MAR mechanism characterised by

$$\Pr(R_Y = 1|R_A = s, Z = z, B = b, Y = y, A = a, L = l) = s,$$
$$\Pr(R_A = 1|Z = z, B = b, Y = y, A = a, L = l) = 0.25 + 0.10b.$$

This mechanism assigns validation data to an individual on either both $Y$ and $A$ (30% of all individuals) or neither depending on their realisation of $B$, the misclassified version of the exposure variable $A$ (Supplementary Table S8.1). Supplementary Tables S8.2 and S8.3 give the likelihood contributions for the parameterisation based on predictive values and the closed form maximum likelihood expressions, respectively. Maximum likelihood estimates can also be found by fitting to the data the saturated logistic regression models of $B$ and $Z$ on $L$ and $(B, L)$, respectively, and to the validation subset the fully saturated logistic regression models of $A$ and $Y$ on $(Z, B, L)$ and $(A, Z, B, L)$, respectively. Estimated weights are then obtained by plugging in the maximum likelihood estimates into (8.8). As in the complete data setting where we assumed the weights to be known, evaluating (8.6) then yields an odds ratio of $\widehat{\text{OR}} = \text{OR} \approx 0.573$.

## 8.4 Simulations

We performed a series of Monte Carlo simulation experiments to illustrate the implementation of the proposed method, to study its finite sample properties and to compare the method to estimators that ignore the presence of confounding or joint exposure and outcome misclassification. All simulations were conducted using R-3.5.0 ((R Core Team, 2018)) on x86_64-pc-linux-gnu platforms of the high performance computer cluster of Leiden University Medical Center.

### *8.4.1 Methods*

For all 54 simulation experiments, we generated $n_{\text{sim}} = 1000$ samples of size $n$ according to the data generating mechanisms depicted in the directed acyclic graphs of Figure 8.1. This multi-step data generating process included generating values on measurement error-free variables, introducing misclassification and allocating individuals validation data. We applied various estimators to each of the simulation samples to yield, for each scenario, an empirical distribution of each point estimator and corresponding precision estimators. These distributions were then summarised into various performance metrics. These metrics include the empirical bias of the estimator on the log-scale (i.e., the mean estimated log-OR minus the target log-OR across the $n_{\text{sim}}$ samples), the empirical standard error (SE) of the estimator on the log-scale (i.e., the square root of the mean squared deviation of the estimated log-OR from the mean log-OR), the empirical mean squared error (MSE) (i.e., the sum of the squared SE and the squared bias), the square root of the mean estimated variance (SSE, sample standard error) and the empirical coverage probability (CP) (i.e., the fraction of simulation runs per scenario where the 95% confidence interval (95%CI) contained the target quantity).

### *1 Distribution of measurement error-free variables*

Following Gravel and Platt (2018), we consider a setting based on that of "Scenario A" in the work of Setoguchi et al. (2008) with slight modifications to the propensity score and outcome models. We consider a fully observed covariate vector $L = (L_0, ..., L_{10})$ whose distribution coincides with that of $h(V)$, where $V = (V_1, ..., V_{10})$ has the multivariate normal distribution with zero means, unit variances and correlations equal to zero except for the correlations between $W_1$ and $V_5$, $V_2$ and $V_6$, $V_3$ and $V_8$, and $V_4$ and $V_9$, which were set to 0.2, 0.9, 0.2, and 0.9, respectively. Function $h$ was defined such that

$$\begin{aligned} h(V) = (&I(V_1 > 0), V_2, I(V_3 > 0), V_4, I(V_5 > 0), I(V_6 > 0), V_7, I(V_8 > 0), \\ &I(V_9 > 0), V_{10}). \end{aligned}$$

Thus, sampling from the distribution of $L$ is equivalent to sampling from the multivariate normal distribution with the given parameter values and dichotomising the 1st, 3rd, 5th, 6th, 8th and 9th elements.

Next, let $U_1$ and $U_2$ be binary variables distributed according to the following logistic models:

$$\operatorname{logit} \Pr(U_1 = 1|L) = \eta_0, \tag{8.10}$$

$$\operatorname{logit} \Pr(U_2 = 1|L, U_1) = \mu_0. \tag{8.11}$$

The distribution of the binary exposure variable $A$ was defined according to the model

$$\operatorname{logit} \Pr(A = 1|L, U_1, U_2) = \alpha_0 + \textstyle\sum_{j=1}^{10} \alpha_j L_j + \alpha_{11} U_1. \tag{8.12}$$

Letting $U_3$ be a scalar random variable that is independent of $(A, L_1, ..., L_{10}, U_1, U_2)$ and uniformly distributed over the interval $[0, 1]$, we defined the counterfactual outcome $Y(a)$, under the intervention setting $A$ to $a$, as

$$Y(a) = I\Big(U_3 < \text{expit}\Big\{\beta_0 + \gamma a + \textstyle\sum_{j=1}^{10} \beta_j L_j + \beta_{11} U_2\Big\}\Big). \tag{8.13}$$

With $Y := Y(A)$, the above implies consistency, conditional exchangeability given $L$ and structural positivity.

## *2 Misclassification mechanism*

For scenarios with joint misclassification, we defined $B = U_1$ and $Z = U_2$, so that the predictive values take a standard logistic form:

$$\text{logit}\,\Pr(Y = 1|A, B, L, Z) = \beta_0 + \gamma A + \textstyle\sum_{j=1}^{10} \beta_j L_j + \beta_{11} Z \tag{8.14}$$

$$\text{logit}\,\Pr(A = 1|B, L, Z) = \alpha_0 + \textstyle\sum_{j=1}^{10} \alpha_j L_j + \alpha_{11} B. \tag{8.15}$$

For scenarios without exposure misclassification, we set $\alpha_{11} = 0$ and defined $B = A$ and $Z = U_2$, so that

$$\text{logit}\,\Pr(Y = 1|A, B, L, Z) = \beta_0 + \gamma A + \textstyle\sum_{j=1}^{10} \beta_j L_j + \beta_{11} Z, \tag{8.16}$$

$$\text{logit}\,\Pr(B = 1|L, Z) = \alpha_0 + \textstyle\sum_{j=1}^{10} \alpha_j L_j. \tag{8.17}$$

For simplicity, we removed any marginal dependence of $Z$ on the covariates $L$ and $U_1$ as well as any marginal dependence of $U_1$ on $L$ (cf. equations (8.10) and (8.11)). Although models (8.10) through (8.15) take a standard logistic form, they do not imply that the corresponding sensitivities and specificities can be written in the same form. We chose the predictive values rather than the sensitivities and specificities to take a standard logistic form so as to ensure correct model specification in the estimation of the weights in the simulation experiments, in which a likelihood approach based on predictive values was adopted (cf. (8.9)).

## *3 Missing data mechanism*

For these simulations, we stipulated $L$, $B$ and $Z$ to be observed for all subjects. We consider scenarios where the dataset can be partitioned into a subset with validation data on all misclassified variables (denoted $R = 1$) and a dataset with validation data on neither ($R = 0$). That is, we simulated data such that subjects have validation data on both $A$ and $Y$ or neither on $A$ nor on $Y$. Values for the response indicator $R$ were generated according to the following (MAR) model:

$$\begin{aligned}\operatorname{logit}\Pr(R=1|Z,B,Y,A,L) &= \operatorname{logit}\Pr(R=1|Z,B,L)\\ &= \xi_0+\xi_1 Z+\xi_2 B+\xi_3 ZB.\end{aligned}$$

## *4 Scenarios*

We initially fixed most parameters of models (8.12) and (8.13) at the respective values of "Scenario A" of Setoguchi et al. (2008): $\alpha_1 = 0.8$, $\alpha_2 = -0.25$, $\alpha_3 = 0.6$, $\alpha_4 = -0.4$, $\alpha_5 = -0.8$, $\alpha_6 = -0.5$, $\alpha_7 = 0.7$, $\alpha_8 = 0$, $\alpha_9 = 0$, $\alpha_{10} = 0$, $\beta_0 = -3.85$, $\beta_1 = 0.3$, $\beta_2 = -0.36$, $\beta_3 = -0.73$, $\beta_4 = -0.2$, $\beta_5 = 0$, $\beta_6 = 0$, $\beta_7 = 0$, $\beta_8 = 0.71$, $\beta_9 = -0.19$ and $\beta_{10} = 0.26$. Parameters $\eta_0$ and $\alpha_0$ were fixed at zero and $\xi_1$, $\xi_2$ and $\xi_3$ at 2, 1 and $-1$, respectively. The remaining parameters and $\beta_0$ were allowed to vary across scenarios as per Table 8.4.

Scenarios differ by sample size $n$, the presence of outcome misclassification, potentially misclassified outcome prevalence (via $\mu_0$), the associations between the exposure and outcome on the one hand and the respective misclassified versions on the other (via $\alpha_{11}$ and $\beta_{11}$), outcome model intercept $\beta_0$, the conditional log-OR $\gamma$, or the size of the validation subset (via $\xi_0$). Based on an iterative Monte Carlo integration approach (Austin and Stafford, 2008), we specified $\gamma$ so as to keep the target marginal log odds ratio at $-0.4$.

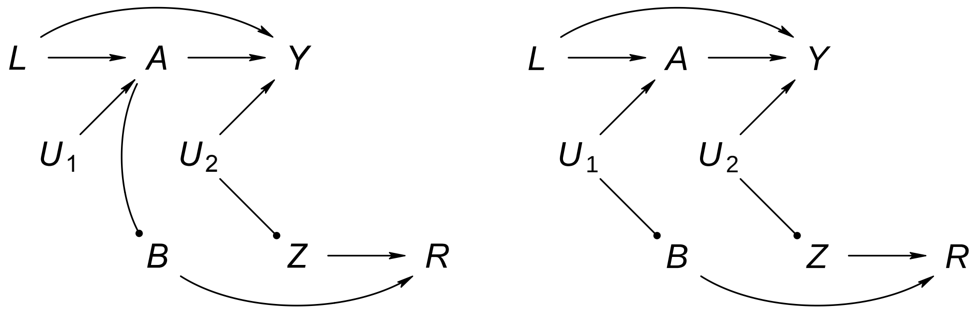


**Figure 8.1:** Data structure for scenarios with misclassification on the outcome only (left) or on both the exposure and outcome (right). Bullet arrowheads represent deterministic relationships.

## 5 *Estimators*

We considered five estimators of the OR for the marginal exposure-outcome effect: a crude estimator (labeled Crude) that ignores both confounding and misclassication of any variable, a misclassification naive estimator (labeled PS) that addresses confounding through IPW, complete cases analysis (CCA) in which IPW is applied only to the subset of subjects with validation data, the Gravel and Platt estimator (GP) that ignores exposure misclassification, and the method proposed in this article (labeled IPWM). Both GP and IPWM are implemented using the R function `mecor::ipwm` (Nab, 2019; Nab et al., 2018), which in the simulation settings considered uses iteratively reweighted least squares via the `stats::glm` function for maximum likelihood estimation. GP coincides with the approach of Gravel and Platt where it concerns point estimation, but they

**Table 8.4:** Simulation parameter values used in the Monte Carlo studies. Scenarios indicated with ‘a’ have $n = 10000$, those with ‘b’ have $n = 5000$ and those with ‘c’ have $n = 1000$.

| Scenarios | Exposure misclassification | $\mu_0$ | $\alpha_{11}$ | $\beta_0$ | $\beta_{11}$ | $\gamma$ | $\xi_0$ |
|---|---|---|---|---|---|---|---|
| 1a,1b,1c | Absent | $-2$ | 0 | $-3.85$ | 2 | $-0.431$ | $-1.5$ |
| 2a,2b,2c | Absent | $-3$ | 0 | $-3.85$ | 2 | $-0.417$ | $-1.5$ |
| 3a,3b,3c | Absent | $-2$ | 0 | $-3.85$ | 4 | $-0.624$ | $-1.5$ |
| 4a,4b,4c | Absent | $-2$ | 0 | $-3.85$ | 2 | $-0.431$ | $-2.5$ |
| 5a,5b,5c | Present | $-2$ | 2 | $-3.85$ | 2 | $-0.431$ | $-1.5$ |
| 6a,6b,6c | Present | $-3$ | 2 | $-3.85$ | 2 | $-0.417$ | $-1.5$ |
| 7a,7b,7c | Present | $-2$ | 4 | $-3.85$ | 2 | $-0.431$ | $-1.5$ |
| 8a,8b,8c | Present | $-2$ | 2 | $-3.85$ | 4 | $-0.624$ | $-1.5$ |
| 9a,9b,9c | Present | $-2$ | 2 | $-3.85$ | 2 | $-0.431$ | $-2.5$ |
| 10a,10b,10c | Absent | $-2$ | 0 | $-2$ | 2 | $-0.470$ | $-1.5$ |
| 11a,11b,11c | Absent | $-3$ | 0 | $-2$ | 2 | $-0.445$ | $-1.5$ |
| 12a,12b,12c | Absent | $-2$ | 0 | $-2$ | 4 | $-0.641$ | $-1.5$ |
| 13a,13b,13c | Absent | $-2$ | 0 | $-2$ | 2 | $-0.470$ | $-2.5$ |
| 14a,14b,14c | Present | $-2$ | 2 | $-2$ | 2 | $-0.470$ | $-1.5$ |
| 15a,15b,15c | Present | $-3$ | 2 | $-2$ | 2 | $-0.445$ | $-1.5$ |
| 16a,16b,16c | Present | $-2$ | 4 | $-2$ | 2 | $-0.470$ | $-1.5$ |
| 17a,17b,17c | Present | $-2$ | 2 | $-2$ | 4 | $-0.641$ | $-1.5$ |
| 18a,18b,18c | Present | $-2$ | 2 | $-2$ | 2 | $-0.470$ | $-2.5$ |

differ in the construction of confidence intervals. Unlike Gravel and Platt (2018), we used a non-parametric rather than a semi-parametric bootstrap procedure for estimating standard errors and constructing confidence intervals. Semi-parametrically generating response indicators would preferably require modelling of (or making additional assumptions about) the missing data mechanism. In particular, to obtain a bootstrap dataset, we defined the record of a unit as their observed data and response indicators, imposed a uniform distribution across all records in the original dataset, and drew independently as many records from this distribution as the total number of records in the original dataset. For all methods and each original dataset, we drew 1000 bootstrap datasets for variance estimation and the construction of percentile confidence intervals.

All estimators are based on a function of the estimated outcome probability $P_1$ in the exposed group and the estimated outcome probability $P_0$ in the unexposed group. However, since $P_1$ and $P_0$ may take a value of 0 or 1, the crude odds ratio $[P_1/(1-P_1)]/[P_0/(1-P_0)]$ need not exist. In contrast to what is often (implicitly) done in simulation studies—i.e., studying the properties of the estimators after conditioning on datasets where $[P_1/(1-P_1)]/[P_0/(1-P_0)]$ is defined—we first define $P_1^* = (P_1 s+1)/(s+2)$ and $P_0^* = (P_0 s+1)/(s+2)$ for a large positive number $s$ (here set to $10^6$) and then regard $[P_1^*/(1-P_1^*)]/[P_0^*/(1-P_0^*)]$ as the estimator of the OR for the exposure-outcome association. This ensures the estimator is always defined and effectively shrinks the outcome probabilities towards 0.5 and the OR towards 1 (Supplementary Appendix S8.2).

For PS and CCA, we used a logistic regression of $B$ and $A$, respectively, on covariates $L_1$ through $L_{10}$ as main effects to estimate the propensity scores. Taking the crude OR for the association between $B$ and $Z$ (PS) or $A$ and $Y$ (CCA) over the data weighted by the reciprocal of the propensity of the received exposure level provided an estimate of the target OR. R code for the methods GP and IPWM is given in Supplementary Appendix S8.3.

### *8.4.2 Results*

The treatment assignment mechanism detailed above resulted in average exposure rates ranging from 17% to 51%, whereas average outcome rates ranged from 3% to 22%. Across all simulation studies, the average outcome rate ranged from 6% to 18%. Across all simulation studies with exposure misclassification, exposure and joint misclassification rates ranged from 16% to 33% and from 2% to 6%, respectively. Approximately 16% to 32% of subjects were allocated validation data.

The results on the performance of the various methods in simulations studies 1-9 are provided in Table 8.5 (see Supplementary Table S8.4 for the results on all scenarios).

As expected, Crude, PS and CCA clearly showed bias with respect to the target log OR of $-0.4$. The bias associated with restricting the analysis to records with validation data is likely brought on to a large extent by collider stratification, with $R$ acting as the collider here (cf. Figure 8.1). Both Crude and PS indicated a null effect, as one would anticipate in view of the marginal and $L$-conditional independence of $B$ and $Z$ implied by the simulation set-up. The empirical coverage probabilities were, although low for both estimators, similar to substantially larger for PS as compared with Crude. Paralleling this is that Crude, whose (implicit) propensity score model is inherently at least as parsimonious, yielded similar to smaller empirical and sample standard errors as compared with PS. With the average fraction of subjects with validation data being as low as 16% (in scenarios with low $\xi_0$) to 32%, it is not unsurprising that Crude was subject to the largest degree of variability.

The results for the IPWM approach are generally favourable for large samples and in line with its theoretical (large sample) properties. For scenarios with smaller samples (scenarios 1c, 2c and 4c, 6c and 9c in particular), however, we observed considerable bias (see Supplementary Table S8.4). Comparing CCA with IPWM, we note a strong linear association between the methods in terms of the absolute within-method differences in estimated bias between scenarios of size 10000 (scenarios labeled ‘a’) and the respective scenarios of size 1000 (scenarios labeled ‘c’) (Pearson correlation 0.997). Note that the results for GP and IPWM are identical for scenarios labeled 1-4 and 10-13 since the methods are equivalent in terms of point estimation in the absence of exposure misclassification. In all other scenarios, i.e., scenarios for which GP was not developed, GP performed substantially worse than IPWM. The non-zero, albeit relatively small, systematic deviations of the IPWM point estimates from the target $-0.4$, notably the estimated bias of $-0.097$ (scenario 2b), may be attributable in part to the outcome being rare (with prevalence ranging from 3% to 8% across scenarios labeled 1-9). This is indicated by the superior performance of IPWM in scenarios where the outcome is more prevalent (cf. scenarios labeled 1-9b versus 10-18b, which have prevalence up to 22%). A similar observation was made by Gravel and Platt (2018).

The standard errors for GP and IPWM were noticeably higher than those of Crude and PS, which is unsurprising in view of the discrepancies in the number of estimated parameters. As expected, increasing the sample size, the true outcome rate (via $\beta_0$) or both led to a decrease in the variability of IPWM (cf. Table 8.4

and Supplementary Table S8.4). However, despite the large discrepancies between SSE and SE for some scenarios, the empirical coverage probabilities of IPWM were close to the nominal level of 0.95, except for scenarios 1c, 2c and 4c, where we observed considerable bias.

## 8.5 Discussion

The analysis of epidemiologic data is often complicated by the presence of confounding and misclassifications in exposure and outcome variables. In this paper we propose a new estimator for estimating a marginal odds-ratio in the presence of confouding and joint misclassification of the exposure and outcome variables. In simulation studies, this weighting estimator showed promising finite sample performance, reducing bias and mean squared error as compared with simpler methods.

The proposed IPWM estimator is an extension of the inverse probability weighting estimator recently proposed by Gravel and Platt (GP) which only addresses the misclassification in the outcome (Gravel and Platt, 2018). IPWM and GP are (mathematically) equivalent when the exposure is (assumed to be) measured without misclassification error.

Like the Gravel and Platt approach, IPWM relies on estimates of sensitivity and specificity or positive and negative predictive values for the misclassified variables. In this paper, we used an internal validation approach where a portion of subjects would receive error-free (‘gold standard’) measurements on either or both the outcome and exposure. However, we anticipate that in some settings the likelihood may not be fully identifiable from the data at hand. In these settings, it may be possible to incorporate external rather than internal information on the misclassification rates, possibly through a Bayesian approach using prior assumptions about misclassification probabilities. When validation data is external, however, it may be necessary to assume misclassification to be independent of covariates $L$, because external studies seldom consider the same covariates as the main study (Lyles et al., 2011). External validation approaches also require the assumption that the misclassification parameters targeted in the validation sample are transportable to the main study.

In the absence of internal and external validation data, it is possible to conduct a sensitivity analysis within the weighting framework. Formula (8.8) for the weights can readily be used in a sensitivity analysis in which the terms describing the distribution of true exposure and outcome variables in relation to the observed data (positive and negative predictive values) serve as sensitivity parameters of

**Table 8.5:** Results for simulation studies 1-9b on the performance of different causal estimators in various scenarios of confounding and misclassification in exposure and outcome. Abbreviations: PS, propensity score method ignoring misclassification; CCA, complete case analysis; GP, Gravel and Platt estimator ignoring exposure misclassification, consistent with the methodology of Gravel and Platt (2018) for point (but not for variance) estimation; IPWM, inverse probability weighting method for confounding and joint exposure and outcome misclassification; BSE, estimated standard error for the bias due to Monte Carlo error; SE, empirical standard error; SSE, sample standard error; CP, empirical coverage probability. In all scenarios, the true marginal log OR (estimand) was $-0.4$.

| | Crude | | | | | |
|---|---|---|---|---|---|---|
| Scenario | Bias | BSE | MSE | SE | SSE | CP |
| 1b | 0.394 | 0.004 | 0.169 | 0.119 | 0.118 | 0.122 |
| 2b | 0.382 | 0.006 | 0.179 | 0.183 | 0.184 | 0.492 |
| 3b | 0.394 | 0.004 | 0.169 | 0.117 | 0.118 | 0.116 |
| 4b | 0.401 | 0.004 | 0.174 | 0.117 | 0.118 | 0.102 |
| 5b | 0.401 | 0.003 | 0.169 | 0.090 | 0.088 | 0.007 |
| 6b | 0.407 | 0.004 | 0.183 | 0.132 | 0.134 | 0.133 |
| 7b | 0.396 | 0.003 | 0.164 | 0.086 | 0.088 | 0.009 |
| 8b | 0.395 | 0.003 | 0.164 | 0.086 | 0.088 | 0.005 |
| 9b | 0.398 | 0.003 | 0.166 | 0.089 | 0.088 | 0.005 |
| | PS | | | | | |
| Scenario | Bias | BSE | MSE | SE | SSE | CP |
| 1b | 0.392 | 0.005 | 0.182 | 0.168 | 0.169 | 0.382 |
| 2b | 0.379 | 0.008 | 0.213 | 0.264 | 0.258 | 0.738 |
| 3b | 0.389 | 0.006 | 0.182 | 0.175 | 0.169 | 0.402 |
| 4b | 0.389 | 0.006 | 0.182 | 0.176 | 0.168 | 0.392 |
| 5b | 0.402 | 0.003 | 0.170 | 0.090 | 0.088 | 0.010 |
| 6b | 0.407 | 0.004 | 0.183 | 0.131 | 0.135 | 0.136 |
| 7b | 0.396 | 0.003 | 0.164 | 0.086 | 0.088 | 0.009 |
| 8b | 0.395 | 0.003 | 0.164 | 0.086 | 0.088 | 0.004 |
| 9b | 0.398 | 0.003 | 0.166 | 0.089 | 0.088 | 0.005 |

the sensitivity analysis. The models for the predictive values can take complex forms, however, thus complicating the analysis and presentation of results.

If internal validation is available, the subjects with validation data need not form a completely random subset. The proposed method, IPWM, was developed under the assumption that validation data allocation occurs in an 'ignorable' fashion (Rubin, 1976). In practice, it may be that the researchers have limited control over the validation data allocation mechanism. For instance, it is conceivable that individuals with specific indications (e.g., with a realisation of $L$, $B$ or $Z$) are practically ineligible to be assigned a double measurement of the exposure ($A$ and $B$) and outcome ($Y$ and $Z$). Further, the estimator also allows for validation subjects to receive either the double exposure or double

Table 8.5 continued.

| | CCA | | | | | |
|---|---|---|---|---|---|---|
| Scenario | Bias | BSE | MSE | SE | SSE | CP |
| 1b | −0.078 | 0.015 | 0.226 | 0.469 | 0.491 | 0.899 |
| 2b | −0.117 | 0.019 | 0.375 | 0.601 | 0.900 | 0.887 |
| 3b | −0.020 | 0.010 | 0.091 | 0.301 | 0.300 | 0.919 |
| 4b | −0.093 | 0.020 | 0.407 | 0.631 | 1.158 | 0.899 |
| 5b | −0.145 | 0.009 | 0.103 | 0.286 | 0.286 | 0.903 |
| 6b | −0.109 | 0.011 | 0.131 | 0.345 | 0.362 | 0.930 |
| 7b | −0.213 | 0.007 | 0.101 | 0.237 | 0.250 | 0.865 |
| 8b | −0.209 | 0.006 | 0.079 | 0.187 | 0.186 | 0.775 |
| 9b | −0.175 | 0.012 | 0.184 | 0.392 | 0.411 | 0.902 |
| | GP | | | | | |
| Scenario | Bias | BSE | MSE | SE | SSE | CP |
| 1b | −0.036 | 0.011 | 0.130 | 0.359 | 0.428 | 0.958 |
| 2b | −0.097 | 0.016 | 0.265 | 0.505 | 0.861 | 0.938 |
| 3b | −0.019 | 0.007 | 0.055 | 0.233 | 0.240 | 0.939 |
| 4b | −0.045 | 0.016 | 0.253 | 0.501 | 1.087 | 0.944 |
| 5b | 0.269 | 0.008 | 0.132 | 0.244 | 0.244 | 0.799 |
| 6b | 0.280 | 0.010 | 0.177 | 0.314 | 0.339 | 0.862 |
| 7b | 0.134 | 0.008 | 0.076 | 0.241 | 0.252 | 0.926 |
| 8b | 0.259 | 0.004 | 0.087 | 0.140 | 0.144 | 0.570 |
| 9b | 0.263 | 0.010 | 0.174 | 0.325 | 0.339 | 0.883 |

Table 8.5 continued.

| | IPWM | | | | | |
|---|---|---|---|---|---|---|
| Scenario | Bias | BSE | MSE | SE | SSE | CP |
| 1b | −0.036 | 0.011 | 0.130 | 0.359 | 0.428 | 0.958 |
| 2b | −0.097 | 0.016 | 0.265 | 0.505 | 0.861 | 0.938 |
| 3b | −0.019 | 0.007 | 0.055 | 0.233 | 0.240 | 0.939 |
| 4b | −0.045 | 0.016 | 0.253 | 0.501 | 1.087 | 0.944 |
| 5b | −0.017 | 0.009 | 0.082 | 0.286 | 0.284 | 0.942 |
| 6b | −0.014 | 0.011 | 0.129 | 0.359 | 0.386 | 0.958 |
| 7b | 0.004 | 0.008 | 0.059 | 0.243 | 0.261 | 0.969 |
| 8b | −0.004 | 0.006 | 0.032 | 0.180 | 0.181 | 0.958 |
| 9b | −0.025 | 0.012 | 0.141 | 0.374 | 0.415 | 0.956 |

outcome measurement. We simulated data such that subjects have validation data on both the exposure and outcome variables or on neither. Although this may greatly simplify analysis and enhance efficiency, in practice it is not necessary to assume that this condition holds. An interesting scenario is where subjects have validation data on at most one variable, i.e., on the exposure variable or the outcome variable but not both. In this case, valid estimation would require additional modelling assumptions; for example, the error-free outcome variable cannot then be regressed on the error-free exposure variable.

To accommodate settings where validation data allocation is not completely at random, we deviated from the semi-parametric bootstrap procedure for variance estimation proposed by Gravel and Platt. Instead, the non-parametric procedure we used requires less assumptions regarding the validation subset sampling procedure. The non-parametric procedure showed good performance in our simulations.

Whilst we have discussed under what conditions the proposed method consistently estimates or at least identifies the target quantity, the assumptions may be untenable in particular settings. Particularly, an infallible measurement tool for the exposure and outcome that can be performed on a subset of the data need not always exist. The robustness to deviations of infallibility is an interesting and important direction for further research. This is especially relevant where there exists considerable uncertainty about the tenability of the assumptions that is difficult to incorporate in the analysis. An obvious and flexible alternative to IPWM is to multiply impute missing values including absent measurement error-free variables before implementing IPW (MI+IPW). Although MI+IPW and IPWM may be comparable in terms of their assumptions, it is yet unclear how they behave under assumption violations such as misspecification of the outcome model.

An advantageous property of MI+IPW is that it can easily accommodate missing covariate values. Other alternatives that can accommodate missing covariates were recently developed by Shu and Yi (2018). Their proposed weighting estimators simultaneously addresses confounding, misclassification of the outcome (but not of the exposure) and measurement error on the covariates under a classical additive measurement error model. The methods can be implemented using validation data or repeated measurements and use a simple misclassification model (in which the outcome surrogate is independent of exposure or covariates given the target outcome) that is suitable for performing sensitivity analyses.

Another interesting area for further research is where the researchers do have control over who is referred for further testing by the assumed infallible measurement tool(s). An obvious choice is to adopt a completely at random strategy (simple random sampling). However, other referral (sampling) strategies exist and it is not clear what strategy leads to the most favourable estimator properties for the given setting.

In summary, we have developed an extension to an existing method, to allow for valid estimation of a marginal causal OR in the presence of confounding and a commonly ignored and misunderstood source of bias—joint exposure and outcome misclassification. The R function `mecor::ipwm` has been made available to facilitate implementation (Nab, 2019; Nab et al., 2018).

## References

Austin, P. C. and J. Stafford (2008): “The performance of two data-generation processes for data with specified marginal treatment odds ratios,” *Communications in Statistics - Simulation and Computation*, 37, 1039–1051.

Babanezhad, M., S. Vansteelandt, and E. Goetghebeur (2010): “Comparison of causal effect estimators under exposure misclassification,” *Journal of Statistical Planning and Inference*, 140, 1306–1319.

Brakenhoff, T. B., M. Mitroiu, R. H. Keogh, K. G. Moons, R. H. Groenwold, and M. van Smeden (2018): “Measurement error is often neglected in medical literature: a systematic review,” *Journal of Clinical Epidemiology*, 98, 89–97.

Braun, D., M. Gorfine, G. Parmigiani, N. D. Arvold, F. Dominici, and C. Zigler (2017): “Propensity scores with misclassified treatment assignment: a likelihood-based adjustment,” *Biostatistics*, 18, 695–710.

Brenner, H., D. A. Savitz, and O. Gefeller (1993): “The effects of joint misclassification of exposure and disease on epidemiologic measures of association exposure and disease on epidemiologic measures of association,” *Journal of Clinical Epidemiology*, 46, 1195–1202.

Brooks, D. R., K. D. Getz, A. T. Brennan, A. Z. Pollack, and M. P. Fox (2018): “The impact of joint misclassification of exposures and outcomes on the results of epidemiologic research,” *Current Epidemiology Reports*, 5, 166–174.

Culver, A. L., I. S. Ockene, R. Balasubramanian, B. C. Olendzki, D. M. Sepavich, J. Wactawski-Wende, J. E. Manson, Y. Qiao, S. Liu, P. A. Merriam, et al. (2012): “Statin use and risk of diabetes mellitus in postmenopausal women in the women’s health initiative,” *Archives of internal medicine*, 172, 144–152.

Dawid, A. (1979): “Conditional independence in statistical theory,” *Journal of the Royal Statistical Society, Series B (Methodological)*, 1–31.

Gravel, C. A. and R. W. Platt (2018): “Weighted estimation for confounded binary outcomes subject to misclassification,” *Statistics in Medicine*, 37, 425–436.

Holland, P. (1986): “Statistics in causal inference,” *Journal of the American Statistical Association*, 81, 945–960.

Holland, P. (1988): “Causal inference, path analysis, and recursive structural equations models,” *Sociological Methodology*, 18, 449–484.

Jurek, A. M., S. Greenland, and G. Maldonado (2008): “Brief report: how far from non-differential does exposure or disease misclassification have to be to bias measures of association away from the null?” *International Journal of Epidemiology*, 37, 382–385.

Jurek, A. M., G. Maldonado, S. Greenland, and T. R. Church (2006): “Exposure-measurement error is frequently ignored when interpreting epidemiologic study results,” *European journal of epidemiology*, 21, 871–876.

Kristensen, P. (1992): “Bias from nondifferential but dependent misclassification of exposure and outcome,” *Epidemiology*, 210–215.

Leong, A., K. Dasgupta, S. Bernatsky, D. Lacaille, A. Avina-Zubieta, and E. Rahme (2013): “Systematic review and meta-analysis of validation studies on a diabetes case definition from health administrative records,” *PloS one*, 8, e75256.

Lyles, R. H., L. Tang, H. M. Superak, C. C. King, D. D. Celentano, Y. Lo, and J. D. Sobel (2011): “Validation data-based adjustments for outcome misclassification in logistic regression: an illustration,” *Epidemiology*, 22, 589.

Marcum, Z. A., M. A. Sevick, and S. M. Handler (2013): “Medication nonadherence: a diagnosable and treatable medical condition,” *Jama*, 309, 2105–2106.

Nab, L. (2019): *mecor: Measurement Error Corrections*, r package version 0.1.0. Available from: https://github.com/LindaNab/mecor.git.

Nab, L., R. H. Groenwold, P. M. Welsing, and M. van Smeden (2018): “Measurement error in continuous endpoints in randomised trials: problems and solutions,” *arXiv preprint arXiv:1809.07068*.

Neyman, J., K. Iwaszkiewicz, and St. Kolodziejczyk (1935): “Statistical problems in agricultural experimentation,” *Supplement to the Journal of the Royal Statistical Society*, 2, 107–180.

Ni, J., A. Leong, K. Dasgupta, and E. Rahme (2017): “Correcting hazard ratio estimates for outcome misclassification using multiple imputation with internal validation data,” *Pharmacoepidemiology and drug safety*, 26, 925–934.

Pearl, J. (2009): *Causality: Models, Reasoning and Inference*, New York: Cambridge University Press.

R Core Team (2018): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `https://www.R-project.org/`.

Rubin, D. (1974): “Estimating causal effects of treatments in randomized and nonrandomized studies,” *Journal of Educational Psychology*, 66, 688–701.

Rubin, D. (1976): “Inference and missing data,” *Biometrika*, 63, 581–592.

Setoguchi, S., S. Schneeweiss, M. B. MA, R. Glynn, and E. Cook (2008): “Evaluating uses of data mining techniques in propensity score estimation: a simulation study,” *Pharmacoepidemiology and Drug Safety*, 17, 546–555.

Shu, D. and G. Y. Yi (2018): “Weighted causal inference methods with mismeasured covariates and misclassified outcomes,” *Statistics in Medicine.*

Shu, D. and G. Y. Yi (2019): “Causal inference with measurement error in outcomes: Bias analysis and estimation methods,” *Statistical Methods in Medical Research*, 28, 2049–2068.

Tang, L., R. H. Lyles, Y. Ye, Y. Lo, and C. C. King (2013): “Extended matrix and inverse matrix methods utilizing internal validation data when both disease and exposure status are misclassified,” *Epidemiologic Methods*, 2, 49–66.

VanderWeele, T. J. and M. A. Hernán (2012): “Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs,” *American journal of epidemiology*, 175, 1303–1310.

Vogel, C., H. Brenner, A. Pfahlberg, and O. Gefeller (2005): “The effects of joint misclassification of exposure and disease on the attributable risk,” *Statistics in Medicine*, 24, 1881–1896.

## Supplementary Material

### S8.1

Suppose $A, B, Y$ and $Z$ are random variables that take values in $\{0, 1\}$.

**Theorem 8.1.** *For any $a, l$, let*

$$\varphi(a,l) = \frac{\varphi^*(a,l)}{\mathbb{E}[\varphi^*(A,L)|A=a]} \quad \textit{and} \quad \varphi^*(a,l) = \frac{1}{\Pr(A=a|L=l)}.$$

*If $Y(A) = Y$ (consistency), $(Y(0), Y(1)) \perp\!\!\!\perp A|L = l$ (conditional exchangeability), $\Pr(A = a) > 0$ and $\Pr(A = a|L = l) > 0$ (positivity) for all $a$ and every $l$ in the support of $L$, then*

$$\mathbb{E}[Y(a)] = \mathbb{E}[\varphi(A,L)I(Y=1)|A=a].$$

*Proof.* We begin by considering $\mathbb{E}[\varphi^*(A,L)|A=a]$. By the law of the unconscious statistician and Bayes' theorem, we have

$$\begin{aligned}
\mathbb{E}[\varphi^*(A,L)|A=a] &= \sum_l \frac{\Pr(L=l|A=a)}{\Pr(A=a|L=l)} \\
&= \sum_l \frac{\Pr(A=a|L=l)\Pr(L=l)}{\Pr(A=a)\Pr(A=a|L=l)} \\
&= \frac{1}{\Pr(A=a)} \sum_l \Pr(L=l) \\
&= \frac{1}{\Pr(A=a)}.
\end{aligned}$$

Hence, for all $a, y$, we have

$$\begin{aligned}
&\sum_l \varphi(a,l)\Pr(Y=y, L=l|A=a) \qquad (8.18)\\
&= \sum_l \frac{\Pr(Y=y,L=l|A=a)\Pr(A=a)}{\Pr(A=a|L=l)} \\
&= \sum_l \frac{\Pr(Y=y|A=a,L=l)\Pr(A=a|L=l)\Pr(L=l)}{\Pr(A=a|L=l)} \\
&= \sum_l \Pr(Y=y|A=a,L=l)\Pr(L=l)
\end{aligned}$$

$$= \sum_l \Pr(Y(a) = y|A = a, L = l) \Pr(L = l) \tag{8.19}$$

$$= \sum_l \Pr(Y(a) = y|L = l) \Pr(L = l) \tag{8.20}$$

$$= \Pr(Y(a) = y),$$

where (8.19) and (8.20) hold under consistency and conditional exchangeability given $L$, respectively. Positivity ensures the weights are defined/exist. Hence, $\mathbb{E}[\varphi(A, L)I(Y = 1)|A = a] = \sum_l \varphi(a, l) \Pr(Y = 1, L = l|A = a) = \mathbb{E}[Y(a)]$, as desired. □

**Corollary 8.1.** *For any $y, a, l$, let*

$$\varphi(a, l) = \frac{\varphi^*(a, l)}{\mathbb{E}[\varphi^*(A, L)|A = a]}, \quad \varphi^*(a, l) = \frac{1}{\Pr(A = a|L = l)}, \quad \textit{and}$$
$$\phi(a, l) = \frac{\Pr(Y = 1, L = l|A = a)}{\Pr(Z = 1, L = l|B = a)}.$$

*If $Y(A) = Y$, $(Y(0), Y(1)) \perp\!\!\!\perp A|L$ and positivity holds, then*

$$\begin{aligned}\mathbb{E}[Y(a)] &= \sum_l \varphi(a, l) \Pr(Y = 1, L = l|A = a)\\ &= \sum_l \varphi(a, l)\phi(a, l) \Pr(Z = 1, L = l|B = a)\\ &= \mathbb{E}[\varphi(B, L)\phi(B, L)I(Z = 1)|B = a].\end{aligned}$$

## S8.2

**Theorem 8.2.** *Fix some $s > 0$ and let $P^* = (Ps + 1)/(s + 2)$ for all $P \in [0, 1]$. If $(P_0, P_1) \in (0, 1) \times (0, 1)$, then*

$$\begin{aligned}1 &< \frac{P_1^*/(1 - P_1^*)}{P_0^*/(1 - P_0^*)} < \frac{P_1/(1 - P_1)}{P_0/(1 - P_0)} \quad \textit{if } P_1 > P_0,\\ 1 &= \frac{P_1^*/(1 - P_1^*)}{P_0^*/(1 - P_0^*)} = \frac{P_1/(1 - P_1)}{P_0/(1 - P_0)} \quad \textit{if } P_1 = P_0, \ \textit{and}\\ 1 &> \frac{P_1^*/(1 - P_1^*)}{P_0^*/(1 - P_0^*)} > \frac{P_1/(1 - P_1)}{P_0/(1 - P_0)} \quad \textit{if } P_1 < P_0\end{aligned}$$

*Proof.* Suppose $(P_0, P_1) \in (0, 1) \times (0, 1)$. If and only if

$$\frac{P_1^*/(1 - P_1^*)}{P_0^*/(1 - P_0^*)} < \frac{P_1/(1 - P_1)}{P_0/(1 - P_0)}, \tag{8.21}$$

then

$$\frac{P_1 s+1}{s+1-P_1 s}\frac{s+1-P_0 s}{P_0 s+1} < \frac{P_1}{1-P_1}\frac{1-P_0}{P_0},$$
$$\frac{P_1 s+1}{s+1-P_1 s}\frac{1-P_1}{P_1} < \frac{P_0 s+1}{s+1-P_0 s}\frac{1-P_0}{P_0}.$$

Now, since

$$\frac{\partial}{\partial P}\left\{\frac{Ps+1}{s+1-Ps}\frac{1-P}{P}\right\} = \frac{(-2P^2+2P-1)S-1}{P^2(1-(P-1)S)^2} < 0$$

over the interval $(0,1)$ for $P$, it follows that inequality (8.21) holds if $P_1 > P_0$. Also, if $P_1 > P_0$, then, since $\partial/(\partial P)\{(Ps+1)/(s+1-Ps)\} > 0$ if $P \in (0,1)$, we have

$$1 < \frac{P_1^*/(1-P_1^*)}{P_0^*/(1-P_0^*)}.$$

Similar arguments establish the assertion for the case where $P_1 < P_0$. It is easily verified that if $P_1 = P_0$, then

$$\frac{P_1^*/(1-P_1^*)}{P_0^*/(1-P_0^*)} = \frac{P_1 s+1}{s+1-P_1 s}\frac{s+1-P_0 s}{P_0 s+1} = 1$$
$$= \frac{P_1}{1-P_1}\frac{1-P_0}{P_0} = \frac{P_1/(1-P_1)}{P_0/(1-P_0)},$$

as desired. □

## S8.3

GP and IPWM were applied to every dataset `data` in R using the function `mecor::ipwm` and the following code:

```
# GP:
formulasGP <- list(
   Y~Z+B+L1+L2+L3+L4+L5+L6+L7+L8+L9+L10,
   B~Z+L1+L2+L3+L4+L5+L6+L7+L8+L9+L10,
   Z~L1+L2+L3+L4+L5+L6+L7+L8+L9+L10
)
mecor::ipwm(
   formulas=formulasGP, data=data, outcome_true=‘‘Y’’,
   outcome_mis=‘‘Z’’, exposure_true=‘‘B’’, exposure_mis=NULL, sp=1e6
)

# IPWM:
formulasIPWM <- list(
   Y~A+Z+B+L1+L2+L3+L4+L5+L6+L7+L8+L9+L10,
   A~Z+B+L1+L2+L3+L4+L5+L6+L7+L8+L9+L10,
   Z~B+L1+L2+L3+L4+L5+L6+L7+L8+L9+L10,
   B~L1+L2+L3+L4+L5+L6+L7+L8+L9+L10
)
mecor::ipwm(
   formulas=formulasIPWM, data=data, outcome_true=‘‘Y’’,
   outcome_mis=‘‘Z’’, exposure_true=‘‘A’’, exposure_mis=‘‘B’’, sp=1e6
)
```

## S8.4 Supplementary Tables

**Table S8.1:** Expected cell counts (rounded to integers) for illustrative study setting after misclassification and formation of validation subsets

| | | | | | $L = 0$ | | $L = 1$ | |
|---|---|---|---|---|---|---|---|---|
| $R_Y$ | $R_A$ | $Y$ | $A$ | $L$ | $A = 0$ | $A = 1$ | $A = 0$ | $A = 1$ |
| 0 | 0 | | | 0 | $m_1 = 9371$ | $m_2 = 7147$ | $m_3 = 1011$ | $m_4 = 884$ |
| 0 | 0 | | | 1 | $m_5 = 1120$ | $m_6 = 3165$ | $m_7 = 80$ | $m_8 = 221$ |
| 0 | 1 | | | | $m_9 = 0$ | $m_{10} = 0$ | $m_{11} = 0$ | $m_{12} = 0$ |
| 1 | 0 | | | | $m_{13} = 0$ | $m_{14} = 0$ | $m_{15} = 0$ | $m_{16} = 0$ |
| 1 | 1 | 0 | 0 | 0 | $m_{17} = 2728$ | $m_{18} = 38$ | $m_{19} = 144$ | $m_{20} = 2$ |
| 1 | 1 | 1 | 0 | 0 | $m_{21} = 13$ | $m_{22} = 3$ | $m_{23} = 169$ | $m_{24} = 53$ |
| 1 | 1 | 0 | 1 | 0 | $m_{25} = 382$ | $m_{26} = 3797$ | $m_{27} = 12$ | $m_{28} = 242$ |
| 1 | 1 | 1 | 1 | 0 | $m_{29} = 1$ | $m_{30} = 9$ | $m_{31} = 12$ | $m_{32} = 178$ |
| 1 | 1 | 0 | 0 | 1 | $m_{33} = 287$ | $m_{34} = 41$ | $m_{35} = 6$ | $m_{36} = 5$ |
| 1 | 1 | 1 | 0 | 1 | $m_{37} = 2$ | $m_{38} = 1$ | $m_{39} = 7$ | $m_{40} = 3$ |
| 1 | 1 | 0 | 1 | 1 | $m_{41} = 84$ | $m_{42} = 1658$ | $m_{43} = 10$ | $m_{44} = 87$ |
| 1 | 1 | 1 | 1 | 1 | $m_{45} = 1$ | $m_{46} = 4$ | $m_{47} = 3$ | $m_{48} = 24$ |

**Table S8.2**: Log-likelihood contributions for all possible types of observations under internal validation sampling.

| Type | $R_Y$ | $R_A$ | $Z$ | $B$ | $Y$ | $A$ | $L$ | Count | Log-likelihood contribution |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | | | 0 | $m_1$ | $\log(1-\varepsilon^*_{00})+\log(1-\delta^*_0)$ |
| 2 | 0 | 0 | 1 | 0 | | | 0 | $m_2$ | $\log(\varepsilon^*_{00})+\log(1-\delta^*_0)$ |
| 3 | 0 | 0 | 0 | 1 | | | 0 | $m_3$ | $\log(1-\varepsilon^*_{10})+\log(\delta^*_0)$ |
| 4 | 0 | 0 | 1 | 1 | | | 0 | $m_4$ | $\log(\varepsilon^*_{10})+\log(\delta^*_0)$ |
| 5 | 0 | 0 | 0 | 0 | | | 1 | $m_5$ | $\log(1-\varepsilon^*_{01})+\log(1-\delta^*_1)$ |
| 6 | 0 | 0 | 1 | 0 | | | 1 | $m_6$ | $\log(\varepsilon^*_{01})+\log(1-\delta^*_1)$ |
| 7 | 0 | 0 | 0 | 1 | | | 1 | $m_7$ | $\log(1-\varepsilon^*_{11})+\log(\delta^*_1)$ |
| 8 | 0 | 0 | 1 | 1 | | | 1 | $m_8$ | $\log(\varepsilon^*_{11})+\log(\delta^*_1)$ |
| 9 | 0 | 1 | | | | | | $m_9+...+m_{12}=0$ | 0 |
| 10 | 1 | 0 | | | | | | $m_{13}+...+m_{16}=0$ | 0 |
| 11 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | $m_{17}$ | $\log(1-\pi^*_{0000})+\log(1-\lambda^*_{000})+\log(1-\varepsilon^*_{00})+\log(1-\delta^*_0)$ |
| 12 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | $m_{18}$ | $\log(1-\pi^*_{0100})+\log(1-\lambda^*_{100})+\log(\varepsilon^*_{00})+\log(1-\delta^*_0)$ |
| 13 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | $m_{19}$ | $\log(1-\pi^*_{0010})+\log(1-\lambda^*_{010})+\log(1-\varepsilon^*_{10})+\log(\delta^*_0)$ |
| 14 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | $m_{20}$ | $\log(1-\pi^*_{0110})+\log(1-\lambda^*_{110})+\log(\varepsilon^*_{10})+\log(\delta^*_0)$ |

Table S8.2 continued.

| Type | $R_Y$ | $R_A$ | $Z$ | $B$ | $Y$ | $A$ | $L$ | Count | Log-likelihood contribution |
|---|---|---|---|---|---|---|---|---|---|
| 15 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | $m_{21}$ | $\log(\pi^*_{0000}) + \log(1 - \lambda^*_{000}) + \log(1 - \varepsilon^*_{00}) + \log(1 - \delta^*_0)$ |
| 16 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | $m_{22}$ | $\log(\pi^*_{0100}) + \log(1 - \lambda^*_{100}) + \log(\varepsilon^*_{00}) + \log(1 - \delta^*_0)$ |
| 17 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | $m_{23}$ | $\log(\pi^*_{0010}) + \log(1 - \lambda^*_{010}) + \log(1 - \varepsilon^*_{10}) + \log(\delta^*_0)$ |
| 18 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | $m_{24}$ | $\log(\pi^*_{0110}) + \log(1 - \lambda^*_{110}) + \log(\varepsilon^*_{10}) + \log(\delta^*_0)$ |
| 19 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | $m_{25}$ | $\log(1 - \pi^*_{1000}) + \log(\lambda^*_{000}) + \log(1 - \varepsilon^*_{00}) + \log(1 - \delta^*_0)$ |
| 20 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | $m_{26}$ | $\log(1 - \pi^*_{1100}) + \log(\lambda^*_{100}) + \log(\varepsilon^*_{00}) + \log(1 - \delta^*_0)$ |
| 21 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | $m_{27}$ | $\log(1 - \pi^*_{1010}) + \log(\lambda^*_{010}) + \log(1 - \varepsilon^*_{10}) + \log(\delta^*_0)$ |
| 22 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | $m_{28}$ | $\log(1 - \pi^*_{1110}) + \log(\lambda^*_{110}) + \log(\varepsilon^*_{10}) + \log(\delta^*_0)$ |
| 23 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | $m_{29}$ | $\log(\pi^*_{1000}) + \log(\lambda^*_{000}) + \log(1 - \varepsilon^*_{00}) + \log(1 - \delta^*_0)$ |
| 24 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | $m_{30}$ | $\log(\pi^*_{1100}) + \log(\lambda^*_{100}) + \log(\varepsilon^*_{00}) + \log(1 - \delta^*_0)$ |
| 25 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | $m_{31}$ | $\log(\pi^*_{1010}) + \log(\lambda^*_{010}) + \log(1 - \varepsilon^*_{10}) + \log(\delta^*_0)$ |
| 26 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | $m_{32}$ | $\log(\pi^*_{1110}) + \log(\lambda^*_{110}) + \log(\varepsilon^*_{10}) + \log(\delta^*_0)$ |

Table S8.2 continued.

| Type | $R_Y$ | $R_A$ | $Z$ | $B$ | $Y$ | $A$ | $L$ | Count | Log-likelihood contribution |
|---|---|---|---|---|---|---|---|---|---|
| 27 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | $m_{33}$ | $\log(1-\pi^*_{0001})+\log(1-\lambda^*_{001})+\log(1-\varepsilon^*_{01})+\log(1-\delta^*_1)$ |
| 28 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | $m_{34}$ | $\log(1-\pi^*_{0101})+\log(1-\lambda^*_{101})+\log(\varepsilon^*_{01})+\log(1-\delta^*_1)$ |
| 29 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | $m_{35}$ | $\log(1-\pi^*_{0011})+\log(1-\lambda^*_{011})+\log(1-\varepsilon^*_{11})+\log(\delta^*_1)$ |
| 30 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | $m_{36}$ | $\log(1-\pi^*_{0111})+\log(1-\lambda^*_{111})+\log(\varepsilon^*_{11})+\log(\delta^*_1)$ |
| 31 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | $m_{37}$ | $\log(\pi^*_{0001})+\log(1-\lambda^*_{001})+\log(1-\varepsilon^*_{01})+\log(1-\delta^*_1)$ |
| 32 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | $m_{38}$ | $\log(\pi^*_{0101})+\log(1-\lambda^*_{101})+\log(\varepsilon^*_{01})+\log(1-\delta^*_1)$ |
| 33 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | $m_{39}$ | $\log(\pi^*_{0011})+\log(1-\lambda^*_{011})+\log(1-\varepsilon^*_{11})+\log(\delta^*_1)$ |
| 34 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | $m_{40}$ | $\log(\pi^*_{0111})+\log(1-\lambda^*_{111})+\log(\varepsilon^*_{11})+\log(\delta^*_1)$ |
| 35 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | $m_{41}$ | $\log(1-\pi^*_{1001})+\log(\lambda^*_{001})+\log(1-\varepsilon^*_{01})+\log(1-\delta^*_1)$ |
| 36 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | $m_{42}$ | $\log(1-\pi^*_{1101})+\log(\lambda^*_{101})+\log(\varepsilon^*_{01})+\log(1-\delta^*_1)$ |
| 37 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | $m_{43}$ | $\log(1-\pi^*_{1011})+\log(\lambda^*_{011})+\log(1-\varepsilon^*_{11})+\log(\delta^*_1)$ |
| 38 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | $m_{44}$ | $\log(1-\pi^*_{1111})+\log(\lambda^*_{111})+\log(\varepsilon^*_{11})+\log(\delta^*_1)$ |

Table S8.2 continued.

| Type | $R_Y$ | $R_A$ | $Z$ | $B$ | $Y$ | $A$ | $L$ | Count | Log-likelihood contribution |
|---|---|---|---|---|---|---|---|---|---|
| 39 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | $m_{45}$ | $\log(\pi^*_{1001}) + \log(\lambda^*_{001}) + \log(1 - \varepsilon^*_{01}) + \log(1 - \delta^*_1)$ |
| 40 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | $m_{46}$ | $\log(\pi^*_{1101}) + \log(\lambda^*_{101}) + \log(\varepsilon^*_{01}) + \log(1 - \delta^*_1)$ |
| 41 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | $m_{47}$ | $\log(\pi^*_{1011}) + \log(\lambda^*_{011}) + \log(1 - \varepsilon^*_{11}) + \log(\delta^*_1)$ |
| 42 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | $m_{48}$ | $\log(\pi^*_{1111}) + \log(\lambda^*_{111}) + \log(\varepsilon^*_{11}) + \log(\delta^*_1)$ |

**Table S8.3**: Closed form expressions for the maximum likelihood estimators (MLE) of the parameters of the likelihood parameterised in terms of predictive values for the hypothetical study setting.

| Parameter | MLE |
|---|---|
| $\delta_0^*$ | $\hat{\delta}_0^* =$ $(m_3+m_4+m_{19}+m_{20}+m_{23}+m_{24}+m_{27}+m_{28}+m_{31}+m_{32})/(\{\sum_{j=1}^{4} m_j\}+\{\sum_{j=17}^{32} m_j\})$ |
| $\delta_1^*$ | $\hat{\delta}_1^* =$ $(m_7+m_8+m_{35}+m_{36}+m_{39}+m_{40}+m_{43}+m_{44}+m_{47}+m_{48})/(\{\sum_{j=5}^{8} m_j\}+\{\sum_{j=33}^{48} m_j\})$ |
| $\varepsilon_{00}^*$ | $\hat{\varepsilon}_{00}^* =$ $(m_2+m_{18}+m_{22}+m_{26}+m_{30})/(m_1+m_2+m_{17}+m_{18}+m_{21}+m_{22}+m_{25}+m_{26}+m_{29}+m_{30})$ |
| $\varepsilon_{10}^*$ | $\hat{\varepsilon}_{10}^* =$ $(m_4+m_{20}+m_{24}+m_{28}+m_{32})/(m_3+m_4+m_{19}+m_{20}+m_{23}+m_{24}+m_{27}+m_{28}+m_{31}+m_{32})$ |
| $\varepsilon_{01}^*$ | $\hat{\varepsilon}_{01}^* =$ $(m_6+m_{34}+m_{38}+m_{42}+m_{46})/(m_5+m_6+m_{33}+m_{34}+m_{37}+m_{38}+m_{41}+m_{42}+m_{45}+m_{46})$ |
| $\varepsilon_{11}^*$ | $\hat{\varepsilon}_{11}^* =$ $(m_8+m_{36}+m_{40}+m_{44}+m_{48})/(m_7+m_8+m_{35}+m_{36}+m_{39}+m_{40}+m_{43}+m_{44}+m_{47}+m_{48})$ |
| $\lambda_{000}^*$ | $\hat{\lambda}_{000}^* = (m_{25} + m_{29})/(m_{17} + m_{21} + m_{25} + m_{29})$ |
| $\lambda_{100}^*$ | $\hat{\lambda}_{100}^* = (m_{26} + m_{30})/(m_{18} + m_{22} + m_{26} + m_{30})$ |
| $\lambda_{010}^*$ | $\hat{\lambda}_{010}^* = (m_{27} + m_{31})/(m_{19} + m_{23} + m_{27} + m_{31})$ |
| $\lambda_{110}^*$ | $\hat{\lambda}_{110}^* = (m_{28} + m_{32})/(m_{20} + m_{24} + m_{28} + m_{32})$ |
| $\lambda_{001}^*$ | $\hat{\lambda}_{001}^* = (m_{41} + m_{45})/(m_{33} + m_{37} + m_{41} + m_{45})$ |
| $\lambda_{101}^*$ | $\hat{\lambda}_{101}^* = (m_{42} + m_{46})/(m_{34} + m_{38} + m_{42} + m_{46})$ |
| $\lambda_{011}^*$ | $\hat{\lambda}_{011}^* = (m_{43} + m_{47})/(m_{35} + m_{39} + m_{43} + m_{47})$ |
| $\lambda_{111}^*$ | $\hat{\lambda}_{111}^* = (m_{44} + m_{48})/(m_{36} + m_{40} + m_{44} + m_{48})$ |

Table S8.3 continued.

| Parameter | MLE |
|---|---|
| $\pi^*_{0000}$ | $\hat{\pi}^*_{0000} = m_{21}/(m_{17} + m_{21})$ |
| $\pi^*_{1000}$ | $\hat{\pi}^*_{1000} = m_{29}/(m_{25} + m_{29})$ |
| $\pi^*_{0100}$ | $\hat{\pi}^*_{0100} = m_{22}/(m_{18} + m_{22})$ |
| $\pi^*_{1100}$ | $\hat{\pi}^*_{1100} = m_{30}/(m_{26} + m_{30})$ |
| $\pi^*_{0010}$ | $\hat{\pi}^*_{0010} = m_{23}/(m_{19} + m_{23})$ |
| $\pi^*_{1010}$ | $\hat{\pi}^*_{1010} = m_{31}/(m_{27} + m_{31})$ |
| $\pi^*_{0110}$ | $\hat{\pi}^*_{0110} = m_{24}/(m_{20} + m_{24})$ |
| $\pi^*_{1110}$ | $\hat{\pi}^*_{1110} = m_{32}/(m_{28} + m_{32})$ |
| $\pi^*_{0001}$ | $\hat{\pi}^*_{0001} = m_{37}/(m_{33} + m_{37})$ |
| $\pi^*_{1001}$ | $\hat{\pi}^*_{1001} = m_{45}/(m_{41} + m_{45})$ |
| $\pi^*_{0101}$ | $\hat{\pi}^*_{0101} = m_{38}/(m_{34} + m_{38})$ |
| $\pi^*_{1101}$ | $\hat{\pi}^*_{1101} = m_{46}/(m_{42} + m_{46})$ |
| $\pi^*_{0011}$ | $\hat{\pi}^*_{0011} = m_{39}/(m_{35} + m_{39})$ |
| $\pi^*_{1011}$ | $\hat{\pi}^*_{1011} = m_{47}/(m_{43} + m_{47})$ |
| $\pi^*_{0111}$ | $\hat{\pi}^*_{0111} = m_{40}/(m_{36} + m_{40})$ |
| $\pi^*_{1111}$ | $\hat{\pi}^*_{1111} = m_{48}/(m_{44} + m_{48})$ |

**Table S8.4:** Results for simulation studies 1a-18a,1b-18b,1c-18c on the performance of different causal estimators in various scenarios of confounding and misclassification in exposure and outcome. Abbreviations: PS, propensity score method ignoring misclassification; CCA, complete case analysis; GP, Gravel and Platt estimator ignoring exposure misclassification; IPWM, inverse probability weighting method for confounding and joint exposure and outcome misclassification; BSE, estimated standard error for the bias due to Monte Carlo error; SE, empirical standard error; SSE, sample standard error; CP, empirical coverage probability. In all scenarios, the true marginal log OR (estimand) was $-0.4$.

| | Crude | | | | | | PS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Bias | BSE | MSE | SE | SSE | CP | Bias | BSE | MSE | SE | SSE | CP |
| 1a | 0.401 | 0.003 | 0.167 | 0.081 | 0.083 | 0.004 | 0.399 | 0.004 | 0.173 | 0.117 | 0.120 | 0.080 |
| 2a | 0.392 | 0.004 | 0.170 | 0.127 | 0.127 | 0.189 | 0.391 | 0.006 | 0.184 | 0.177 | 0.181 | 0.436 |
| 3a | 0.400 | 0.003 | 0.167 | 0.083 | 0.083 | 0.005 | 0.391 | 0.004 | 0.167 | 0.119 | 0.119 | 0.104 |
| 4a | 0.394 | 0.003 | 0.162 | 0.081 | 0.083 | 0.007 | 0.392 | 0.004 | 0.169 | 0.122 | 0.120 | 0.106 |
| 5a | 0.398 | 0.002 | 0.162 | 0.061 | 0.062 | 0.000 | 0.398 | 0.002 | 0.162 | 0.061 | 0.062 | 0.000 |
| 6a | 0.404 | 0.003 | 0.172 | 0.094 | 0.094 | 0.010 | 0.404 | 0.003 | 0.172 | 0.094 | 0.095 | 0.011 |
| 7a | 0.399 | 0.002 | 0.163 | 0.062 | 0.062 | 0.000 | 0.399 | 0.002 | 0.163 | 0.062 | 0.062 | 0.000 |
| 8a | 0.401 | 0.002 | 0.165 | 0.064 | 0.062 | 0.000 | 0.401 | 0.002 | 0.165 | 0.064 | 0.062 | 0.000 |
| 9a | 0.400 | 0.002 | 0.164 | 0.064 | 0.062 | 0.000 | 0.400 | 0.002 | 0.164 | 0.064 | 0.062 | 0.000 |
| 10a | 0.396 | 0.003 | 0.164 | 0.085 | 0.083 | 0.004 | 0.395 | 0.004 | 0.171 | 0.123 | 0.119 | 0.101 |
| 11a | 0.396 | 0.004 | 0.173 | 0.128 | 0.127 | 0.176 | 0.388 | 0.006 | 0.185 | 0.187 | 0.182 | 0.455 |
| 12a | 0.398 | 0.003 | 0.165 | 0.081 | 0.083 | 0.007 | 0.398 | 0.004 | 0.173 | 0.120 | 0.120 | 0.096 |
| 13a | 0.399 | 0.003 | 0.166 | 0.083 | 0.083 | 0.004 | 0.395 | 0.004 | 0.171 | 0.120 | 0.119 | 0.102 |
| 14a | 0.404 | 0.002 | 0.167 | 0.061 | 0.062 | 0.000 | 0.404 | 0.002 | 0.167 | 0.061 | 0.062 | 0.000 |
| 15a | 0.398 | 0.003 | 0.167 | 0.092 | 0.094 | 0.011 | 0.398 | 0.003 | 0.167 | 0.092 | 0.095 | 0.012 |
| 16a | 0.404 | 0.002 | 0.167 | 0.063 | 0.062 | 0.000 | 0.404 | 0.002 | 0.167 | 0.063 | 0.062 | 0.000 |
| 17a | 0.399 | 0.002 | 0.163 | 0.061 | 0.062 | 0.000 | 0.399 | 0.002 | 0.163 | 0.061 | 0.062 | 0.000 |
| 18a | 0.401 | 0.002 | 0.164 | 0.059 | 0.062 | 0.000 | 0.401 | 0.002 | 0.164 | 0.059 | 0.062 | 0.000 |

Table S8.4 continued.

| | Crude | | | | | | PS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Bias | BSE | MSE | SE | SSE | CP | Bias | BSE | MSE | SE | SSE | CP |
| 1b | 0.394 | 0.004 | 0.169 | 0.119 | 0.118 | 0.122 | 0.392 | 0.005 | 0.182 | 0.168 | 0.169 | 0.382 |
| 2b | 0.382 | 0.006 | 0.179 | 0.183 | 0.184 | 0.492 | 0.379 | 0.008 | 0.213 | 0.264 | 0.258 | 0.738 |
| 3b | 0.394 | 0.004 | 0.169 | 0.117 | 0.118 | 0.116 | 0.389 | 0.006 | 0.182 | 0.175 | 0.169 | 0.402 |
| 4b | 0.401 | 0.004 | 0.174 | 0.117 | 0.118 | 0.102 | 0.389 | 0.006 | 0.182 | 0.176 | 0.168 | 0.392 |
| 5b | 0.401 | 0.003 | 0.169 | 0.090 | 0.088 | 0.007 | 0.402 | 0.003 | 0.170 | 0.090 | 0.088 | 0.010 |
| 6b | 0.407 | 0.004 | 0.183 | 0.132 | 0.134 | 0.133 | 0.407 | 0.004 | 0.183 | 0.131 | 0.135 | 0.136 |
| 7b | 0.396 | 0.003 | 0.164 | 0.086 | 0.088 | 0.009 | 0.396 | 0.003 | 0.164 | 0.086 | 0.088 | 0.009 |
| 8b | 0.395 | 0.003 | 0.164 | 0.086 | 0.088 | 0.005 | 0.395 | 0.003 | 0.164 | 0.086 | 0.088 | 0.004 |
| 9b | 0.398 | 0.003 | 0.166 | 0.089 | 0.088 | 0.005 | 0.398 | 0.003 | 0.166 | 0.089 | 0.088 | 0.005 |
| 10b | 0.397 | 0.004 | 0.171 | 0.117 | 0.118 | 0.100 | 0.396 | 0.005 | 0.185 | 0.167 | 0.170 | 0.387 |
| 11b | 0.391 | 0.006 | 0.185 | 0.179 | 0.183 | 0.466 | 0.362 | 0.008 | 0.199 | 0.261 | 0.253 | 0.732 |
| 12b | 0.401 | 0.004 | 0.174 | 0.118 | 0.118 | 0.109 | 0.391 | 0.005 | 0.182 | 0.173 | 0.169 | 0.394 |
| 13b | 0.404 | 0.004 | 0.176 | 0.111 | 0.117 | 0.080 | 0.396 | 0.005 | 0.185 | 0.169 | 0.167 | 0.367 |
| 14b | 0.400 | 0.003 | 0.168 | 0.087 | 0.088 | 0.008 | 0.400 | 0.003 | 0.168 | 0.087 | 0.088 | 0.006 |
| 15b | 0.397 | 0.004 | 0.176 | 0.135 | 0.134 | 0.161 | 0.397 | 0.004 | 0.176 | 0.135 | 0.135 | 0.161 |
| 16b | 0.401 | 0.003 | 0.168 | 0.087 | 0.088 | 0.006 | 0.400 | 0.003 | 0.168 | 0.087 | 0.088 | 0.006 |
| 17b | 0.403 | 0.003 | 0.170 | 0.087 | 0.088 | 0.003 | 0.403 | 0.003 | 0.170 | 0.087 | 0.088 | 0.004 |
| 18b | 0.400 | 0.003 | 0.168 | 0.087 | 0.088 | 0.004 | 0.400 | 0.003 | 0.168 | 0.088 | 0.088 | 0.003 |

Table S8.4 continued.

| | Crude | | | | | | PS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Bias | BSE | MSE | SE | SSE | CP | Bias | BSE | MSE | SE | SSE | CP |
| 1c | 0.394 | 0.009 | 0.232 | 0.277 | 0.275 | 0.698 | 0.366 | 0.013 | 0.292 | 0.398 | 0.391 | 0.871 |
| 2c | 0.334 | 0.018 | 0.423 | 0.558 | 0.844 | 0.873 | 0.256 | 0.022 | 0.563 | 0.706 | 0.924 | 0.916 |
| 3c | 0.383 | 0.009 | 0.222 | 0.274 | 0.276 | 0.739 | 0.371 | 0.013 | 0.297 | 0.399 | 0.393 | 0.875 |
| 4c | 0.375 | 0.009 | 0.218 | 0.278 | 0.277 | 0.732 | 0.332 | 0.013 | 0.276 | 0.407 | 0.392 | 0.880 |
| 5c | 0.405 | 0.006 | 0.204 | 0.200 | 0.199 | 0.470 | 0.405 | 0.006 | 0.205 | 0.201 | 0.199 | 0.474 |
| 6c | 0.410 | 0.010 | 0.261 | 0.304 | 0.317 | 0.724 | 0.410 | 0.010 | 0.263 | 0.308 | 0.318 | 0.729 |
| 7c | 0.406 | 0.006 | 0.203 | 0.196 | 0.199 | 0.469 | 0.406 | 0.006 | 0.204 | 0.198 | 0.200 | 0.469 |
| 8c | 0.404 | 0.006 | 0.204 | 0.202 | 0.199 | 0.474 | 0.405 | 0.006 | 0.205 | 0.201 | 0.200 | 0.470 |
| 9c | 0.406 | 0.006 | 0.202 | 0.192 | 0.198 | 0.468 | 0.404 | 0.006 | 0.201 | 0.193 | 0.199 | 0.470 |
| 10c | 0.384 | 0.009 | 0.222 | 0.272 | 0.276 | 0.717 | 0.359 | 0.013 | 0.288 | 0.399 | 0.388 | 0.873 |
| 11c | 0.358 | 0.014 | 0.324 | 0.443 | 0.825 | 0.864 | 0.296 | 0.020 | 0.471 | 0.619 | 0.902 | 0.923 |
| 12c | 0.377 | 0.008 | 0.212 | 0.265 | 0.277 | 0.749 | 0.343 | 0.013 | 0.284 | 0.407 | 0.393 | 0.878 |
| 13c | 0.377 | 0.008 | 0.210 | 0.259 | 0.276 | 0.741 | 0.341 | 0.013 | 0.274 | 0.397 | 0.390 | 0.888 |
| 14c | 0.411 | 0.006 | 0.206 | 0.192 | 0.199 | 0.446 | 0.411 | 0.006 | 0.206 | 0.192 | 0.200 | 0.458 |
| 15c | 0.393 | 0.009 | 0.241 | 0.294 | 0.315 | 0.764 | 0.393 | 0.009 | 0.242 | 0.296 | 0.316 | 0.770 |
| 16c | 0.399 | 0.006 | 0.198 | 0.196 | 0.198 | 0.484 | 0.399 | 0.006 | 0.198 | 0.196 | 0.200 | 0.482 |
| 17c | 0.395 | 0.006 | 0.193 | 0.191 | 0.199 | 0.471 | 0.394 | 0.006 | 0.191 | 0.190 | 0.199 | 0.474 |
| 18c | 0.402 | 0.006 | 0.201 | 0.197 | 0.199 | 0.478 | 0.403 | 0.006 | 0.202 | 0.199 | 0.200 | 0.482 |

Table S8.4 continued.

| | CCA | | | | | | GP | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Bias | BSE | MSE | SE | SSE | CP | Bias | BSE | MSE | SE | SSE | CP |
| 1a | −0.011 | 0.010 | 0.091 | 0.302 | 0.315 | 0.932 | −0.016 | 0.008 | 0.059 | 0.242 | 0.257 | 0.960 |
| 2a | −0.038 | 0.013 | 0.165 | 0.404 | 0.398 | 0.909 | −0.024 | 0.010 | 0.108 | 0.328 | 0.353 | 0.956 |
| 3a | 0.004 | 0.007 | 0.044 | 0.210 | 0.208 | 0.939 | −0.013 | 0.005 | 0.028 | 0.167 | 0.165 | 0.943 |
| 4a | −0.050 | 0.014 | 0.189 | 0.432 | 0.441 | 0.905 | −0.022 | 0.011 | 0.116 | 0.341 | 0.371 | 0.944 |
| 5a | −0.128 | 0.006 | 0.054 | 0.194 | 0.199 | 0.890 | 0.271 | 0.005 | 0.100 | 0.163 | 0.168 | 0.633 |
| 6a | −0.097 | 0.007 | 0.066 | 0.237 | 0.245 | 0.926 | 0.269 | 0.007 | 0.121 | 0.221 | 0.225 | 0.772 |
| 7a | −0.232 | 0.005 | 0.082 | 0.168 | 0.173 | 0.736 | 0.118 | 0.005 | 0.043 | 0.171 | 0.173 | 0.904 |
| 8a | −0.197 | 0.004 | 0.056 | 0.130 | 0.130 | 0.646 | 0.263 | 0.003 | 0.079 | 0.098 | 0.101 | 0.261 |
| 9a | −0.173 | 0.008 | 0.101 | 0.266 | 0.270 | 0.883 | 0.257 | 0.007 | 0.116 | 0.224 | 0.229 | 0.795 |
| 10a | 0.017 | 0.005 | 0.029 | 0.169 | 0.170 | 0.953 | 0.003 | 0.005 | 0.022 | 0.147 | 0.152 | 0.946 |
| 11a | 0.007 | 0.006 | 0.040 | 0.200 | 0.193 | 0.947 | −0.014 | 0.006 | 0.039 | 0.196 | 0.203 | 0.952 |
| 12a | 0.058 | 0.005 | 0.028 | 0.157 | 0.154 | 0.928 | −0.003 | 0.004 | 0.019 | 0.138 | 0.136 | 0.943 |
| 13a | 0.018 | 0.007 | 0.056 | 0.236 | 0.237 | 0.946 | −0.003 | 0.006 | 0.037 | 0.192 | 0.194 | 0.940 |
| 14a | −0.092 | 0.003 | 0.018 | 0.099 | 0.105 | 0.864 | 0.265 | 0.003 | 0.079 | 0.091 | 0.095 | 0.191 |
| 15a | −0.051 | 0.004 | 0.016 | 0.115 | 0.119 | 0.933 | 0.264 | 0.004 | 0.084 | 0.121 | 0.124 | 0.421 |
| 16a | −0.166 | 0.003 | 0.036 | 0.094 | 0.092 | 0.559 | 0.138 | 0.003 | 0.028 | 0.096 | 0.096 | 0.710 |
| 17a | −0.116 | 0.003 | 0.023 | 0.095 | 0.094 | 0.762 | 0.264 | 0.003 | 0.076 | 0.080 | 0.082 | 0.110 |
| 18a | −0.115 | 0.005 | 0.035 | 0.149 | 0.144 | 0.859 | 0.266 | 0.004 | 0.088 | 0.131 | 0.128 | 0.455 |
| 1b | −0.078 | 0.015 | 0.226 | 0.469 | 0.491 | 0.899 | −0.036 | 0.011 | 0.130 | 0.359 | 0.428 | 0.958 |
| 2b | −0.117 | 0.019 | 0.375 | 0.601 | 0.900 | 0.887 | −0.097 | 0.016 | 0.265 | 0.505 | 0.861 | 0.938 |
| 3b | −0.020 | 0.010 | 0.091 | 0.301 | 0.300 | 0.919 | −0.019 | 0.007 | 0.055 | 0.233 | 0.240 | 0.939 |
| 4b | −0.093 | 0.020 | 0.407 | 0.631 | 1.158 | 0.899 | −0.045 | 0.016 | 0.253 | 0.501 | 1.087 | 0.944 |
| 5b | −0.145 | 0.009 | 0.103 | 0.286 | 0.286 | 0.903 | 0.269 | 0.008 | 0.132 | 0.244 | 0.244 | 0.799 |
| 6b | −0.109 | 0.011 | 0.131 | 0.345 | 0.362 | 0.930 | 0.280 | 0.010 | 0.177 | 0.314 | 0.339 | 0.862 |
| 7b | −0.213 | 0.007 | 0.101 | 0.237 | 0.250 | 0.865 | 0.134 | 0.008 | 0.076 | 0.241 | 0.252 | 0.926 |
| 8b | −0.209 | 0.006 | 0.079 | 0.187 | 0.186 | 0.775 | 0.259 | 0.004 | 0.087 | 0.140 | 0.144 | 0.570 |
| 9b | −0.175 | 0.012 | 0.184 | 0.392 | 0.411 | 0.902 | 0.263 | 0.010 | 0.174 | 0.325 | 0.339 | 0.883 |

Table S8.4 continued.

| | CCA | | | | | | GP | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Bias | BSE | MSE | SE | SSE | CP | Bias | BSE | MSE | SE | SSE | CP |
| 10b | 0.011 | 0.007 | 0.056 | 0.237 | 0.244 | 0.957 | −0.002 | 0.007 | 0.050 | 0.223 | 0.221 | 0.939 |
| 11b | 0.001 | 0.009 | 0.083 | 0.288 | 0.276 | 0.918 | −0.019 | 0.010 | 0.093 | 0.304 | 0.304 | 0.938 |
| 12b | 0.058 | 0.007 | 0.050 | 0.216 | 0.223 | 0.953 | −0.007 | 0.006 | 0.038 | 0.194 | 0.197 | 0.949 |
| 13b | −0.015 | 0.011 | 0.122 | 0.350 | 0.345 | 0.934 | −0.023 | 0.009 | 0.077 | 0.277 | 0.287 | 0.950 |
| 14b | −0.092 | 0.005 | 0.030 | 0.146 | 0.148 | 0.889 | 0.263 | 0.004 | 0.088 | 0.136 | 0.136 | 0.505 |
| 15b | −0.060 | 0.005 | 0.033 | 0.170 | 0.170 | 0.929 | 0.263 | 0.006 | 0.101 | 0.177 | 0.183 | 0.712 |
| 16b | −0.171 | 0.004 | 0.047 | 0.132 | 0.131 | 0.741 | 0.139 | 0.004 | 0.038 | 0.136 | 0.138 | 0.820 |
| 17b | −0.121 | 0.004 | 0.032 | 0.134 | 0.135 | 0.842 | 0.263 | 0.004 | 0.082 | 0.115 | 0.118 | 0.388 |
| 18b | −0.113 | 0.007 | 0.055 | 0.206 | 0.207 | 0.904 | 0.264 | 0.006 | 0.102 | 0.178 | 0.185 | 0.702 |
| 1c | −1.163 | 0.098 | 10.972 | 3.101 | 3.092 | 0.792 | −0.994 | 0.095 | 10.003 | 3.003 | 3.085 | 0.878 |
| 2c | −2.086 | 0.131 | 21.614 | 4.155 | 3.415 | 0.733 | −1.979 | 0.130 | 20.875 | 4.118 | 3.689 | 0.835 |
| 3c | −0.184 | 0.029 | 0.880 | 0.920 | 1.477 | 0.887 | −0.102 | 0.024 | 0.574 | 0.751 | 1.412 | 0.939 |
| 4c | −2.730 | 0.148 | 29.275 | 4.671 | 3.710 | 0.722 | −2.295 | 0.142 | 25.436 | 4.491 | 3.974 | 0.916 |
| 5c | −0.254 | 0.029 | 0.904 | 0.916 | 1.764 | 0.891 | 0.288 | 0.018 | 0.409 | 0.571 | 1.106 | 0.951 |
| 6c | −0.548 | 0.067 | 4.832 | 2.129 | 2.684 | 0.893 | 0.402 | 0.046 | 2.276 | 1.454 | 2.782 | 0.969 |
| 7c | −0.223 | 0.020 | 0.467 | 0.646 | 1.303 | 0.912 | 0.109 | 0.020 | 0.424 | 0.642 | 1.322 | 0.952 |
| 8c | −0.236 | 0.014 | 0.258 | 0.450 | 0.508 | 0.891 | 0.279 | 0.011 | 0.197 | 0.345 | 0.369 | 0.883 |
| 9c | −0.662 | 0.079 | 6.624 | 2.487 | 3.186 | 0.896 | 0.314 | 0.053 | 2.915 | 1.678 | 2.454 | 0.972 |
| 10c | −0.105 | 0.019 | 0.376 | 0.604 | 0.740 | 0.897 | −0.098 | 0.017 | 0.294 | 0.534 | 0.790 | 0.943 |
| 11c | −0.113 | 0.025 | 0.649 | 0.798 | 1.174 | 0.906 | −0.183 | 0.031 | 1.006 | 0.986 | 1.758 | 0.931 |
| 12c | 0.010 | 0.018 | 0.313 | 0.560 | 0.598 | 0.938 | −0.053 | 0.021 | 0.455 | 0.673 | 0.703 | 0.937 |
| 13c | −0.170 | 0.036 | 1.330 | 1.140 | 1.967 | 0.919 | −0.129 | 0.030 | 0.920 | 0.950 | 1.983 | 0.948 |
| 14c | −0.107 | 0.011 | 0.133 | 0.349 | 0.360 | 0.922 | 0.286 | 0.010 | 0.187 | 0.324 | 0.351 | 0.882 |
| 15c | −0.088 | 0.013 | 0.187 | 0.424 | 0.417 | 0.913 | 0.276 | 0.015 | 0.314 | 0.488 | 0.664 | 0.951 |
| 16c | −0.159 | 0.009 | 0.115 | 0.299 | 0.311 | 0.924 | 0.143 | 0.010 | 0.124 | 0.323 | 0.357 | 0.952 |
| 17c | −0.133 | 0.010 | 0.114 | 0.311 | 0.323 | 0.918 | 0.259 | 0.009 | 0.144 | 0.277 | 0.304 | 0.872 |
| 18c | −0.148 | 0.016 | 0.264 | 0.492 | 0.593 | 0.938 | 0.280 | 0.014 | 0.278 | 0.447 | 0.510 | 0.926 |

Table S8.4 continued.

| | IPWM | | | | | |
|---|---|---|---|---|---|---|
| Scenario | Bias | BSE | MSE | SE | SSE | CP |
| 1a | −0.016 | 0.008 | 0.059 | 0.242 | 0.257 | 0.960 |
| 2a | −0.024 | 0.010 | 0.108 | 0.328 | 0.353 | 0.956 |
| 3a | −0.013 | 0.005 | 0.028 | 0.167 | 0.165 | 0.943 |
| 4a | −0.022 | 0.011 | 0.116 | 0.341 | 0.371 | 0.944 |
| 5a | −0.004 | 0.006 | 0.035 | 0.186 | 0.194 | 0.954 |
| 6a | −0.010 | 0.008 | 0.060 | 0.244 | 0.252 | 0.952 |
| 7a | −0.013 | 0.005 | 0.030 | 0.172 | 0.179 | 0.951 |
| 8a | 0.003 | 0.004 | 0.015 | 0.122 | 0.125 | 0.952 |
| 9a | −0.019 | 0.008 | 0.065 | 0.255 | 0.265 | 0.948 |
| 10a | 0.003 | 0.005 | 0.022 | 0.147 | 0.152 | 0.946 |
| 11a | −0.014 | 0.006 | 0.039 | 0.196 | 0.203 | 0.952 |
| 12a | −0.003 | 0.004 | 0.019 | 0.138 | 0.136 | 0.943 |
| 13a | −0.003 | 0.006 | 0.037 | 0.192 | 0.194 | 0.940 |
| 14a | −0.005 | 0.003 | 0.011 | 0.104 | 0.106 | 0.962 |
| 15a | −0.001 | 0.004 | 0.017 | 0.129 | 0.134 | 0.963 |
| 16a | 0.010 | 0.003 | 0.010 | 0.099 | 0.099 | 0.947 |
| 17a | 0.001 | 0.003 | 0.009 | 0.096 | 0.095 | 0.943 |
| 18a | 0.001 | 0.005 | 0.022 | 0.148 | 0.144 | 0.949 |
| 1b | −0.036 | 0.011 | 0.130 | 0.359 | 0.428 | 0.958 |
| 2b | −0.097 | 0.016 | 0.265 | 0.505 | 0.861 | 0.938 |
| 3b | −0.019 | 0.007 | 0.055 | 0.233 | 0.240 | 0.939 |
| 4b | −0.045 | 0.016 | 0.253 | 0.501 | 1.087 | 0.944 |
| 5b | −0.017 | 0.009 | 0.082 | 0.286 | 0.284 | 0.942 |
| 6b | −0.014 | 0.011 | 0.129 | 0.359 | 0.386 | 0.958 |
| 7b | 0.004 | 0.008 | 0.059 | 0.243 | 0.261 | 0.969 |
| 8b | −0.004 | 0.006 | 0.032 | 0.180 | 0.181 | 0.958 |
| 9b | −0.025 | 0.012 | 0.141 | 0.374 | 0.415 | 0.956 |

Table S8.4 continued.

| | IPWM | | | | | |
|---|---|---|---|---|---|---|
| Scenario | Bias | BSE | MSE | SE | SSE | CP |
| 10b | −0.002 | 0.007 | 0.050 | 0.223 | 0.221 | 0.939 |
| 11b | −0.019 | 0.010 | 0.093 | 0.304 | 0.304 | 0.938 |
| 12b | −0.007 | 0.006 | 0.038 | 0.194 | 0.197 | 0.949 |
| 13b | −0.023 | 0.009 | 0.077 | 0.277 | 0.287 | 0.950 |
| 14b | −0.003 | 0.005 | 0.022 | 0.147 | 0.152 | 0.960 |
| 15b | −0.006 | 0.006 | 0.035 | 0.187 | 0.198 | 0.963 |
| 16b | 0.010 | 0.004 | 0.020 | 0.142 | 0.143 | 0.956 |
| 17b | −0.003 | 0.004 | 0.017 | 0.131 | 0.136 | 0.956 |
| 18b | 0.010 | 0.006 | 0.042 | 0.205 | 0.207 | 0.955 |
| 1c | −0.994 | 0.095 | 10.003 | 3.003 | 3.085 | 0.878 |
| 2c | −1.979 | 0.130 | 20.875 | 4.118 | 3.689 | 0.835 |
| 3c | −0.102 | 0.024 | 0.574 | 0.751 | 1.412 | 0.939 |
| 4c | −2.295 | 0.142 | 25.436 | 4.491 | 3.974 | 0.916 |
| 5c | −0.101 | 0.029 | 0.849 | 0.916 | 1.771 | 0.950 |
| 6c | −0.373 | 0.069 | 4.896 | 2.181 | 3.041 | 0.978 |
| 7c | −0.027 | 0.022 | 0.470 | 0.685 | 1.298 | 0.961 |
| 8c | −0.019 | 0.014 | 0.200 | 0.447 | 0.527 | 0.953 |
| 9c | −0.372 | 0.068 | 4.769 | 2.152 | 2.579 | 0.989 |
| 10c | −0.098 | 0.017 | 0.294 | 0.534 | 0.790 | 0.943 |
| 11c | −0.183 | 0.031 | 1.006 | 0.986 | 1.758 | 0.931 |
| 12c | −0.053 | 0.021 | 0.455 | 0.673 | 0.703 | 0.937 |
| 13c | −0.129 | 0.030 | 0.920 | 0.950 | 1.983 | 0.948 |
| 14c | −0.003 | 0.011 | 0.130 | 0.360 | 0.396 | 0.967 |
| 15c | −0.018 | 0.016 | 0.263 | 0.512 | 0.705 | 0.984 |
| 16c | 0.014 | 0.011 | 0.114 | 0.338 | 0.371 | 0.966 |
| 17c | −0.005 | 0.010 | 0.101 | 0.318 | 0.349 | 0.952 |
| 18c | −0.002 | 0.016 | 0.249 | 0.499 | 0.613 | 0.962 |

# 9

# Negative controls: concepts and caveats

Bas B. L. Penning de Vries
Rolf H. H. Groenwold

[Submitted]

Negative controls

## Abstract

Unmeasured confounding is a well-known obstacle in causal inference. In recent years, negative controls have received increasing attention as a important tool to address concerns about the problem. The literature on the topic has expanded rapidly and several authors have advocated the more routine use of negative controls in epidemiological practice. In this paper, we review concepts and methodologies based on negative controls for detection and correction of unmeasured confounding bias. We argue that negative controls may lack both specificity and sensitivity to detect unmeasured confounding and that proving the null hypothesis of a null negative control association is impossible. We focus our discussion on the control outcome calibration approach, the difference-in-difference approach, and the double-negative control approach as methods for confounding correction. For each of these methods, we highlight their assumptions and illustrate the potential impact of violations thereof. Given the potentially large impact of assumption violations, it may sometimes be desirable to replace strong conditions for exact identification with weaker, easily verifiable conditions, even when these imply at most partial identification of unmeasured confounding. Future research in this area may broaden the applicability of negative controls and in turn make them better suited for routine use in epidemiological practice. At present, however, the applicability of negative controls should be carefully judged on a case-by-case basis.

Chapter 9

## 9.1 Introduction

In epidemiological research on causal effects, there are often concerns that one or more assumptions—such as exchangeability, no measurement error, or assumptions about missing data—are violated. In efforts to lessen these concerns, it has long been suggested that auxiliary variables be used that have a known (e.g., null) causal relation with the exposure or outcome of interest (Rosenbaum, 1989; Lipsitch et al., 2010; Flanders et al., 2011). Observing an association that contradicts the belief in a causal null might alert the analyst to violations of the assumptions underlying the methods used in the study. Auxiliary variables known to be causally unrelated to the variables of primary interest are called negative controls and have potential in bias detection as well as partial or complete bias correction in epidemiological research (Shi et al., 2020b).

In recent years, negative controls have received increasing attention in the epidemiological and statistical literature. The literature on how to leverage negative controls in studies on causal effects has rapidly expanded and several authors have argued that negative controls should be more commonly employed (Lipsitch et al., 2010; Arnold et al., 2016; Shi et al., 2020b). This paper aims to complement these efforts to increase the more routine implementation of negative controls with a discussion about a selection of caveats. Focusing on the use of negative controls to address possible violations of the exchangeability assumption, i.e., the assumption of no unmeasured confounding, we begin with a brief review of relevant definitions and discuss assumptions for bias detection. We then review methods for bias correction and study their sensitivity to assumption violations.

## 9.2 Negative controls

A negative control outcome (NCO) is a variable that is not causally affected by the exposure of interest $A$ (Tchetgen Tchetgen, 2013; Shi et al., 2020b). Likewise, a negative control exposure (NCE) is a variable that does not causally affect the outcome of interest $Y$, except possibly through the exposure of interest (Shi et al., 2020b). The causal DAGs of Figure 9.1 (discussed later in this section) give examples of settings where a variable $Z$ classifies as an NCO, an NCE or both. Given the absence of a direct causal effect of exposure $A$ on an NCO $Z$ or of NCE $Z$ on outcome $Y$, any observed association between $A$ and an $Z$, or between an $Z$ and outcome $Y$ given $A$, must be spurious. Leveraging negative controls involves translating information about such spurious associations into information about the spuriousness of associations between the primary exposure and outcome variables of interest.

### 9.2.1 Negative controls for unmeasured confounding detection

Let $Y(a)$ denote the outcome that would be realised had exposure $A$ been set to $a$. Together with causal consistency (i.e., $Y(a) = Y$ if $A = a$) and positivity, epidemiologists often seek to invoke the exchangeability (or unmeasured confounding) condition $Y(a) \perp\!\!\!\perp A$ (possibly within levels of a collection of observed variables) to establish identifiability of the effect of exposure $A$ on outcome $Y$ (Hernán and Robins, 2020). In observational studies, however, it is seldom evident that the exchangeability condition, E, for the exposure-outcome relation of interest is achieved. A key idea of negative controls is to find a 'control' statement, C, that translates into information about E and which is more easily verified or refuted.

Control statement C may refer to the absence of bias of a measure of the association between $A$ and $Y$ and the NCO or NCE variable, respectively. Knowing that any control association is noncausal renders the control statement empirically verifiable. If C implies E, then a null finding for the control statement would imply conditional exchangeability for the exposure-outcome relation of interest. Conversely, if E implies C, evidence of bias of the control association corroborates the existence of unmeasured confounding.

### 9.2.2 Caveats in the use of negative controls to detect unmeasured confounding

There are a number of caveats concerning the use of negative controls for confounding detection. These caveats mainly concern the link between the control statement and exchangeability for the exposure-outcome relation of interest. Unfortunately, the extent to which one confers information about the other need

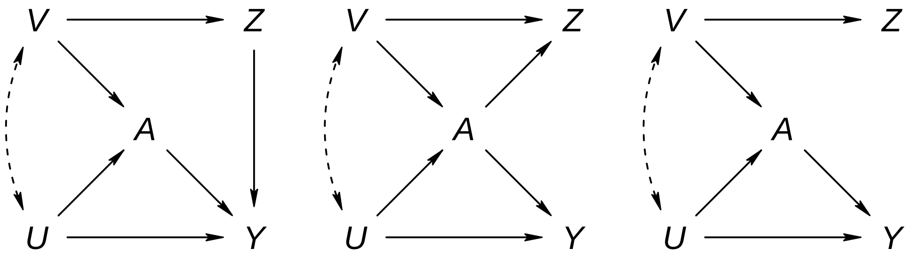


**Figure 9.1:** Causal directed acyclic graphs of settings where $Z$ is a negative control outcome (left), a negative control exposure (middle) or both (right).

not be evident (Groenwold, 2013). A biased negative-control association need not imply unmeasured confounding for the exposure-outcome relation of interest and neither is the converse true generally.

First, while most applications of negative controls assume that confounding is the only source of bias, in reality it may be one of potentially many sources of bias. A spurious negative control association could have resulted, at least in part, from collider stratification, measurement error or violations of assumptions about missing data (Arnold et al., 2016). Even if unmeasured confounding for the negative control association implies unmeasured confounding for the exposure-outcome relation of interest, a biased negative control association need not be a reflection of unmeasured confounding. Conversely, a (near) null finding could be the result of opposing biases, masking the presence of unmeasured confounding. In other words, negative controls are a tool that may lack both specificity and sensitivity with respect to the type(s) of bias they are to detect.

Lipsitch et al. (2010) suggested a principle for establishing a link that is based on the extent to which common causes of $A$ and $Y$ overlap with the common causes of the exposure or outcome and the negative control variable. Clearly, for an NCO, with complete overlap (e.g., $V = U$ in Figure 9.1), the set of common causes of $A$ and $Y$ is empty if and only if the set of common causes of $A$ and the NCO is empty. However, null values for certain measures of the effect of $A$ on an NCO or of an NCE on $Y$ need not imply that the set of unobserved common causes is empty, or, therefore, that there is conditional exchangeability for the primary exposure-outcome relation. Indeed, near null values may be the result of partially opposing confounding effects (or, more generally, opposing biases) and the relative effects may be different for the NCO versus the primary outcome $Y$.

With finite samples rather than complete knowledge of the theoretical or population distribution, sampling variability becomes relevant too, making it more important to acknowledge the distinction between absence of evidence and evidence of absence (Albert and Anderson, 1984). With finite samples, proving the null hypothesis of a null negative control association is impossible. Even if ‘highly’ powered studies cannot detect bias for the negative control relation, it may be injudicious to assume that the available data are sufficient to adequately control for confounding of the primary relation of interest, because a small degree of bias for the former relation may be associated with a substantial degree of bias for the latter. Sample size and power considerations are often ignored or left at secondary importance. While some papers have considered the power of negative control tests (Rosenbaum, 1989; Birch, 1964), it is typically ignored how the negative control association relates to the extent of bias for the exposure-outcome relation of interest, yet high power to detect ‘small departures’ from exposure-

NCO or NCE-outcome independence need not imply high power to detect small bias due to unmeasured confounding of the primary relation of interest. What are considered ‘small departures’ should therefore depend on the relation between the negative control association and the bias for the exposure-outcome relation of interest. Conversely, even if there is evidence of the contrary to the negative control null hypothesis, the bias due to uncontrolled confounding for the primary exposure-outcome relation may not be meaningful. In any case, it is important to consider the relative size of the biases in the negative control and primary exposure-outcome relations.

## 9.3 Negative control methods for uncontrolled confounding adjustment

The more recent literature on negative controls has considered how and under what conditions negative controls can be leveraged to partially or fully identify target causal quantities rather than merely the presence of bias. Lipsitch et al. (2010) gives conditions for valid inference about the direction of bias and thus for partial identification of the target causal quantity. These conditions are reviewed in Supplementary Appendix S9.1. In what follows, we review three methods for full identification: the control outcome calibration approach (COCA), the (generalised) difference-in-difference approach, and the double-negative control approach. Proofs of identification are given in Supplementary Appendix S9.2 for completeness. For each of the methods, we illustrate the potential impact of assumption violations on the identifiability of the targeted quantity. Throughout, departures from identification are termed bias.

### *9.3.1 Control outcome calibration approach*

*Identification*

It may be tempting to regard the confounded association between the exposure of interest and an NCO as a direct measure of bias for the exposure-outcome effect of interest. However, it cannot generally be assumed that the direction or magnitude of bias are the same for the two relations. As an alternative to the restrictive and probably unrealistic “bias equivalence” assumption, i.e., the assumption of equality between between the confounded negative control association and the bias due to unmeasured confounding of the exposure-outcome effect of interest, Tchetgen Tchetgen (2013) proposed the COCA. The assumption of “bias equivalence” would especially likely be violated if the NCO and primary outcome are measured on different scales and the bias is bounded differently depending on the scale, such as would be the case if the NCO were binary and the primary outcome continuous. The COCA leverages an NCO to adjust for unmeasured confounding without requiring that the NCO and primary outcome are measured on similar scales.

The next result, due to Tchetgen Tchetgen (2013), describes a regression-based approach to implementing the COCA, which—characteristically of the COCA—relies on the assumption that a (set of) counterfactual primary outcome(s) of interest is sufficient to render the NCO conditionally independent of the exposure of interest. Some intuition behind this approach may be obtained upon noting that the counterfactual outcomes may well capture information about baseline covariates and therefore serve as a proxy for unobserved pre-exposure variables that are predictive of the NCO. The reasoning rests on the assumption that the same covariates that explain the lack of exchangeability for the outcome of interest also explain the confounding of the exposure-NCO relation. However, even then it is not evident nor guaranteed that the counterfactual outcome proxy is sufficient to render the NCO and exposure conditionally independent.

**Theorem 9.1** (A regression-based approach to implementing the COCA under rank preservation)**.** *Suppose that the following conditions hold for all levels $a$ of $A$:*

- Consistency: $Y(a) = Y$ *if* $a = A$.
- Rank preservation: *for some constant* $\theta$, $Y(0) = Y(a) - \theta a$.
- Exposure-NCO independence given counterfactual outcome: $Z \perp\!\!\!\perp A | Y(0)$.
- NCO model: *for known one-to-one model link* $g$,

$g(\mathbb{E}[Z|A, Y]) = \beta_0 + \beta_1 A + \beta_2 Y$, *where* $\beta_0, \beta_1, \beta_2$ *are identified by a regression of* $Z$ *on* $A$ *and* $Y$, *and* $\beta_2 \neq 0$.

*Then,* $\mathbb{E}[Y(a) - Y(a-1)] = \theta$ *is identified by* $-\beta_1/\beta_2$.

Because counterfactual outcome $Y(0)$ may not fully account for the unmeasured confounding between the exposure and NCO, it is important that the impact of assumption violations be gauged. To this end, Tchetgen Tchetgen (2013) described a sensitivity analysis, given below in Theorem 9.2, for the special case of Theorem 9.1 where $g$ is the identity link and $A$ is a linear combination of $Y(0)$ and an error term $\Delta$. When the sensitivity parameter ($\rho$) is set to 0, it is implicitly assumed that the NCO and exposure of interest are independent given counterfactual outcome $Y(0)$ (because $\chi$ is independent of $(A, Y)$ and therefore of $Y(0)$) and, so, the result of Theorem 9.1 is recovered.

**Theorem 9.2** (Sensitivity analysis for violations of $Z \perp\!\!\!\perp A|Y(0)$)**.** *Suppose the following conditions hold for all levels $a$ of $A$:*

- Consistency: $Y(a) = Y$ *if* $a = A$.
- Rank preservation: *for some constant* $\theta$*,* $Y(0) = Y(a) - \theta a$.
- Conditional exposure-NCO independence: $Z \perp\!\!\!\perp A|(Y(0), \Delta)$.
- Exposure model: $A = \alpha_0 + \alpha_1 Y(0) + \Delta$.
- NCO model: $Z = \beta_0 + \beta_1 Y(0) + \rho\Delta + \chi$, $\chi \perp\!\!\!\perp (A, Y)$.

*Then,* $\mathbb{E}[Z|A, Y] = \beta_0^* + \beta_1^* A + \beta_2^* Y$ *for some* $\beta_0^*, \beta_1^*, \beta_2^*$*, and if parameters* $\beta_1^*, \beta_2^*$ *are identified (by a regression of* $Z$ *on* $A$ *and* $Y$*) and* $\beta_2^* \neq 0$*, then* $\theta = (\beta_1^* - \rho)/\beta_2^*$.

Through the rank preservation assumption, Theorem 9.1 relies also on the strong assumption that the set of all counterfactual outcomes of an individual are deterministically linked. A prerequisite of this assumption is that the within-person ranks of counterfactuals are the same for all individuals. In the next section, we consider violations of this assumption. However, as Theorem 9.3 states, in the special case where the outcome and exposure of interest are binary, there should be no concern about violations of this assumption as it can be dropped entirely (Tchetgen Tchetgen, 2013).

**Theorem 9.3** (COCA for binary primary outcome and exposure)**.** *Suppose that the following conditions hold:*

- Consistency: $Y(a) = Y$ *if* $a = A$
- Positivity: $0 < \Pr(A = a, Y = y)$ *for* $y = 0, 1$.
- Exposure-NCO independence given counterfactual outcome: $Z \perp\!\!\!\perp A|Y(a)$.
- Non-zero denominator: $\mathbb{E}[Z|A = a, Y = 1] - \mathbb{E}[Z|A = a, Y = 0] \neq 0$.

*Then,*

$$\begin{aligned}\mathbb{E}[Y(a)] = \mathbb{E}[Y|A = a]\Pr(A = a)& \\ + \frac{\mathbb{E}[Z|A = 1 - a] - \mathbb{E}[Z|A = a, Y = 0]}{\mathbb{E}[Z|A = a, Y = 1] - \mathbb{E}[Z|A = a, Y = 0]}&\Pr(A = 1 - a).\end{aligned}$$

If the assumptions of Theorem 9.3 are met for $a = 1$, the average treatment effect among the treated (ATT) $\mathbb{E}[Y - Y(0)|A = 1]$ is identified. For identification of the average treatment effect (ATE) $\mathbb{E}[Y(1) - Y(0)]$, the result requires that the assumptions are met for $a = 0, 1$. We will consider violations of these assumptions in the next section.

*Sensitivity to assumption violations*

In this subsection, we consider the sensitivity of the COCA to assumption violations. In particular we illustrate the potential impact of deviating from rank preservation and of violating the assumption that counterfactual outcome $Y(0)$ renders the exposure and NCO conditionally independent. While classical measurement error in the outcome does not hamper inference in terms of bias in the classical linear regression setting, we also illustrate that this for of measurement error does result in bias of the COCA.

First, to illustrate the potential impact of deviating from rank preservation, consider the setting where $A$ is binary and where the following models hold:

$$\left.\begin{aligned} \theta|A &\sim \text{Normal}(\mathbb{E}[\theta], \sigma_\theta^2), \\ Y(0)|A,\theta &\sim \text{Normal}(\alpha_0 + \alpha_1 A, \sigma_Y^2), \\ Y = Y(A) &= Y(0) + \theta A, \\ Z|(A,\theta,Y(0)) &\sim \text{Normal}(\gamma_0 + \gamma_1 Y(0), \sigma_Z^2). \end{aligned}\right\} \tag{9.1}$$

A standard implementation of the COCA as per Theorem 9.1 yields $\hat{\theta} = -\hat{\beta}_1/\hat{\beta}_2$, where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the coefficients for $A$ and $Y$ of an ordinary least squares regression of $Z$ on $A$ and $Y$.

Given a value of the ATE (i.e, $\mathbb{E}[\theta]$), the parameter values are fully determined under models (9.1) by the joint distribution of the observed variables $A, Y, Z$ (Supplementary Appendix S9.3). In particular, given a fixed distribution of $(A, Y, Z)$, the variance of the individual effects $Y(1) - Y(0)$ (i.e., $\text{Var}(\theta) = \sigma_\theta^2$) and the ATE are linearly related via

$$\text{Var}(\theta) = \frac{\text{Var}(A)\text{Var}(Y) - \text{Cov}(A,Y)^2}{(\text{Var}(A) + \mathbb{E}[A]^2)\text{Cov}(A,Z)}(\hat{\beta}_1 - \hat{\beta}_2\mathbb{E}[\theta])$$

(Supplementary Appendix S9.3). For values of the ATE between $-4$ and 2, we chose parameter values such that the distribution of $(A, Y, Z)$ has marginal means $\mathbb{E}[A] = 0.25$, $\mathbb{E}[Y] = 0$ and $\mathbb{E}[Z] = 0$, and covariance matrix

$$\begin{bmatrix} 3/16 & 1/2 & 1/2 \\ 1/2 & 3 & 2 \\ 1/2 & 2 & 4 \end{bmatrix}. \tag{9.2}$$

Figure 9.2 shows the bias of the COCA for the ATE. As shown, the magnitude of bias is zero under rank preservation but increases linearly with increasing variance of individual exposure-outcome effects.

In illustrating the sensitivity of the COCA against violations of rank preservation, it was assumed that the other assumptions were maintained. We now turn to the assumption of Exposure-NCO independence given counterfactual outcome $Y(0)$ and likewise assume that all other assumptions, including rank preservation, are met. In particular, we consider the setting where $Y(0)$ is the sum of two independent variables $U_1, U_2$. By assuming the following models, we also stipulate that some (albeit not necessarily the same) linear combination $\alpha_0' + \alpha_1' U_1 + \alpha_2' U_2$ is sufficient to render the exposure of interest and NCO conditionally independent:

$$\left.\begin{aligned} U_1 &\perp\!\!\!\perp U2, \\ A|(U_1, U_2) &\sim \text{Normal}(\alpha_0 + \alpha_1 U_1 + \alpha_2 U_2, \sigma_A^2), \\ Y = Y(A) &= U_1 + U_2 + \theta A,\ \theta \text{ constant}, \\ Z|(U_1, U_2, A, Y) &\sim \text{Normal}(\alpha_0' + \alpha_1' U_1 + \alpha_2' U_2, \sigma_Z^2) \end{aligned}\right\} \tag{9.3}$$

Variables $U_1$ and $U_2$ can be viewed as common causes of the NCO and the exposure and outcome of interest. Again, the COCA identifies the quantity $\hat{\theta} = -\hat{\beta}_1/\hat{\beta}_2$ based on an ordinary least squares regression of NCO $Z$ on $A$ and $Y$, but this quantity is not generally equal to $\theta$. Figure 9.3 shows the asymptotic bias (departure from identification of the ATE) of the COCA plotted against $\alpha_2$ over the interval $(-5, 5)$ for the special case where $U_1$ and $U_2$ take the standard

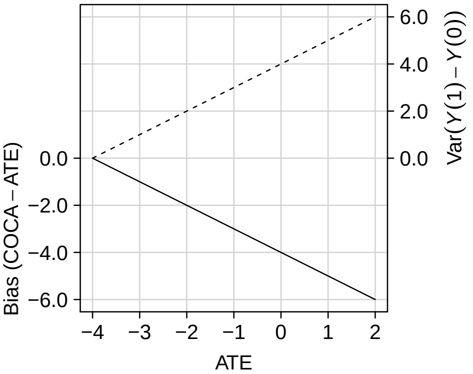


**Figure 9.2:** Illustration of the effect of violating the rank preservation assumption on the difference between the quantity identified by the COCA and the ATE (bias). The dashed line depicts the relation between the variance of individual exposure outcome effects $Y(1) - Y(0)$ and the mean $\mathbb{E}[Y(1) - Y(0)]$ (the ATE) under a fixed observed data distribution; the solid line describes the relation between the ATE and the bias of the implementation of the COCA.

normal distribution and where $\alpha_0, \alpha_0', \alpha_2' = 0$, $\alpha_1, \sigma_A^2, \sigma_Z^2 = 1$ and $\alpha_1' = 2$. The bias is zero only when counterfactual outcome $Y(0)$ is proportional to the linear combination of common causes $U_1$ and $U_2$ that renders the NCO and exposure of interest conditionally independent.

With $\alpha_2, \alpha_2' = 0$, models (9.3) imply the same joint distribution of observed variables $A, Y, Z$ as models (9.4):

$$\left.\begin{aligned} U_1 &\perp\!\!\!\perp U2, \\ A|(U_1, U_2) &\sim \text{Normal}(\alpha_0 + \alpha_1 U_1, \sigma_A^2), \\ Y(A) &= U_1 + \theta A,\ \theta \text{ constant}, \\ Y &= Y(A) + U_2, \\ Z|(U_1, U_2, A, Y) &\sim \text{Normal}(\alpha_0' + \alpha_1' U_1, \sigma_Z^2) \end{aligned}\right\} \tag{9.4}$$

An important difference between (9.3) and (9.4) is that the consistency assumption is violated (provided that $\text{Var}(U_2) > 0$). The observed outcome $Y$ is now the sum of the outcome of interest $Y(A)$ and an independent mean-zero error term. Figure 9.3 therefore also illustrates that the validity of the COCA also critically rests on the absence of classical measurement error in the outcome. At $\alpha_2 = 0$, Figure 9.3 gives the bias of the COCA under (9.4) with the values for the parameters given above. Although ATE $\theta$ may not be identified in the presence of classical measurement error, in Supplementary Appendix S9.3, partial identification bounds are derived for $\theta$.

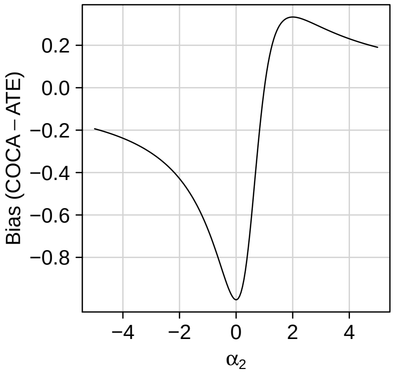


**Figure 9.3:** Illustration of the potential impact of violating the the assumption that the NCO and exposure of interest are independent given counterfactual outcome $Y(0)$.

### *9.3.2 Difference-in-difference approach*

#### *Identification*

The difference-in-difference approach (DiD) proposed by Sofer et al. (2016) is an alternative approach to the COCA and does not assume rank preservation, nor does it require that the counterfactual outcome $Y(0)$ renders the NCO and exposure of interest conditionally independent. Instead, the approach relies on bias equivalence for the primary exposure-outcome relation and the exposure-NCO relation. The simplest version of the DiD approach identifies the ATT under additive equi-confounding, as stated in Theorem 9.4, via the difference between the crude difference in primary outcome means and the bias of the exposure-NCO relation.

**Theorem 9.4** (Difference-in-difference approach for the ATT under additive equi-confounding)**.** *Suppose that the following conditions hold for all levels* $a = 0, 1$*:*
- Consistency: $Y(a) = Y$ *if* $a = A$*.*
- Additive equi-confounding:

$\mathbb{E}[Y(0)|A = 1] - \mathbb{E}[Y(0)|A = 0] = \mathbb{E}[N|A = 1] - \mathbb{E}[N|A = 0]$.

*Then,* $\mathbb{E}[Y(1) - Y(0)|A = 1] = (\mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0]) - (\mathbb{E}[N|A = 1] - \mathbb{E}[N|A = 0])$.

Additive equi-confounding is relatively easy to interpret. However, the assumption may be particularly likely to be violated when primary outcome $Y$ and NCO $Z$ are measured on different scales. A generalized DiD approach still identifies the ATT under a different constraint on the dependence between $Y(0)$ and $A$ in relation to the dependence between $N$ and $A$. In particular, Theorem 9.5, based on Sofer et al. (2016), relies on quantile-quantile equi-confounding, an example of which is depicted in Figure 9.4.

**Theorem 9.5** (Generalized difference-in-difference approach for the ATT under quantile-qualine equi-confounding)**.** *Suppose that the following conditions hold for all levels* $a = 0, 1$*:*
- Consistency: $Y(a) = Y$ *if* $a = A$*.*
- Quantile-quantile equi-confounding: $F_0(F_1^{-1}(p)) = G_0(G_1^{-1}(p))$ *for all* $p \in [0, 1]$*, where* $F_a(y) = \Pr(Y(0) \leq y|A = a)$*,* $F_a^{-1}(p) = \min\{y : p \leq F_a(y)\}$*,* $G_a(z) = \Pr(Z \leq z|A = a)$*,* $G_a^{-1}(p) = \min\{z : p \leq G_a(z)\}$*.*
- $F_1$ *is strictly increasing.*

*Then,* $\mathbb{E}[Y(1) - Y(0)|A = 1] = \mathbb{E}[Y|A = 1] - \mathbb{E}[F_0^{-1}(G_0(G_1^{-1}(V)))]$*, where* $V \sim \text{Uniform}[0, 1]$*.*

*Sensitivity to assumption violations*

We now give a simple setting where neither additive nor quantile-quantile equi-confounding is guaranteed to hold. The setting is characterised by two common causes $U_1, U_2$ of the primary exposure and outcome and of the NCO. As before, we let the relative effects of these common causes to differ between exposure, primary outcome and NCO, and we suppose that the following models hold:

$$\left.\begin{aligned} A &\sim \text{Bernoulli}(p_A), \\ U_1|A &\sim \text{Normal}(\alpha_0 + \alpha_1 A, \sigma_1^2), \\ U_2|(U_1, A) &\sim \text{Normal}(\alpha_0' + \alpha_1' A, \sigma_2^2), \\ Y(0)|(U_1, U_2, A) &\sim \text{Normal}(U_1 + U_2, \sigma_Y^2), \\ Y = Y(A) &= Y(0) + \theta A, \ \theta \text{ constant}, \\ Z|(U_1, U_2, A, Y(0)) &\sim \text{Normal}(\beta_0 + \beta_1 U_1 + \beta_2 U_2, \sigma_Z^2). \end{aligned}\right\} \tag{9.5}$$

Parameters $\alpha_1, \alpha_1', \beta_1, \beta_2$ control the dependence (confounding), through $U_1$ and $U_2$, between $A$ and $Y(0)$ and between $A$ and NCO $Z$; in the special case where these parameters take the value 0, there is no confounding. The models of (9.5) imply

$$Y(0)|A \sim \text{Normal}((\alpha_0 + \alpha_0') + (\alpha_1 + \alpha_1')A, \sigma_1^2 + \sigma_2^2 + \sigma_Y^2),$$

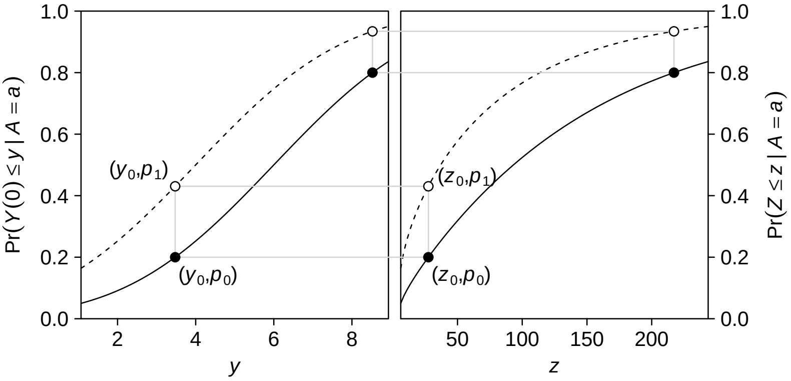


**Figure 9.4:** Example of quantile-quantile equi-confounding. Dashed curves represents $a = 1$, solid curves $a = 0$. There is quantile-quantile equi-confounding because for every two points $(y_0, p_0)$ and $(y_0, p_1)$ on the solid and dashed curves, respectively, of the left panel, there exists $z_0$ such that $(z_0, p_0)$ and $(z_0, p_1)$ lie on the solid and dashed curves, respectively, of the right panel; quantiles $y_0$ and $z_0$ need not be the same.

$$\begin{aligned}&Y|A \sim \text{Normal}((\alpha_0+\alpha_0')+(\alpha_1+\alpha_1'+\theta)A, \sigma_1^2+\sigma_2^2+\sigma_Y^2),\\&Z|A \sim \text{Normal}((\beta_0+\beta_1\alpha_0+\beta_2\alpha_0')+(\beta_1\alpha_1+\beta_2\alpha_1')A), \beta_1^2\sigma_1^2+\beta_2^2\sigma_2^2+\sigma_Z^2).\end{aligned}$$

Implementing the DiD for the ATT $\theta$ would therefore identify, under (9.5), the quantity

$$\begin{aligned}&(\mathbb{E}[Y|A=1]-\mathbb{E}[Y|A=0])-(\mathbb{E}[N|A=1]-\mathbb{E}[N|A=0])\\&\qquad=(1-\beta_1)\alpha_1+(1-\beta_2)\alpha_1'+\theta,\end{aligned}$$

with a bias of $(1-\beta_1)\alpha_1+(1-\beta_2)\alpha_1'$. The generalised DiD would instead identify

$$\begin{aligned}&\mathbb{E}[Y|A=1]-\mathbb{E}[F_0^{-1}(G_0(G_1^{-1}(V)))]\\&\qquad=(\alpha_0+\alpha_0')+(\alpha_1+\alpha_1'+\theta)-\int_{-\infty}^{+\infty}F_0^{-1}(G_0(G_1^{-1}(p)))\,\mathrm{d}p,\end{aligned}$$

where $G_1^{-1}$ is the quantile function of the associated with the distribution of $Z|A=1$, $G_0$ is the cumulative distribution function for $Z|A=1$ and $F_0^{-1}$ the quantile function of $Y|A=0$.

Figure 9.5 shows, for various parameter specifications, the bias of the (generalised) DiD for the ATT $\theta$. Specifically, $\beta_1$ was varied over $(-2,2)$ and $\alpha_1'$ over $\{0,1\}$, while $\beta_2$ was set to $2-\beta_1$, and $p_A=0.5$, $\alpha_0,\alpha_0',\beta_0,\theta=0$ and $\alpha_1=1,\sigma_1^2,\sigma_2^2,,\sigma_Y^2,\sigma_Z^2=1$. The figure illustrates that under additive and quantile-quantile equi-confounding the DiD and generalised DiD, respectively, identify the ATT. It also shows that both approaches are sensitive—albeit differently—to violations of their respective assumptions. Interestingly, even in the absence of additive equi-confounding the generalised DiD could be subject to considerable bias (Figure 9.5, right panel, where the bias for the DiD is $(1-\beta_1)\alpha_1+(1-\beta_2)\alpha_2'=2-(\beta_1+\beta_2)=0$). Beside the interpretability of its assumptions, an appealing property of the standard DiD approach is that the effects of common causes need not be the same for the NCO and primary outcome of interest; if the net additive confounding is (close to) the same for the NCO and primary outcome, then the ATT may be (nearly) identified.

### *9.3.3 Double-negative control approach*

#### *Identification*

Recent developments on the use of negative controls to adjust for unmeasured confounding leverage multiple negative control variables or proxies of unmeasured common causes (Miao et al., 2018a,b; Shi et al., 2020a,b; Tchetgen et al., 2020).

For example, the next result, due to Miao et al. (2018b), gives a set of conditions sufficient to identify the expected marginal counterfactual outcome $\mathbb{E}[Y(a)]$ by leveraging a pair of proxy variables $B, Z$ of an unobserved variable $U$ that renders the counterfactual outcomes independent of the exposure of interest (i.e., conditional exchangeability given $U$).

**Theorem 9.6** (The confounding bridge approach)**.** *Suppose that for all levels $a$ of $A$, the following conditions hold:*
- Consistency: $Y(a) = Y$ *if* $a = A$.
- Positivity: $0 < \Pr(A = a|B) < 1$ *with probability 1.*
- Latent ignorability: $Y(a) \perp\!\!\!\perp (A, B)|U$ *and* $Z \perp\!\!\!\perp (A, B)|U$.
- Confounding bridge assumption: $\mathbb{E}[Y|A = a, U] = \mathbb{E}[h(Z)|A = a, U]$ *with probability 1 for some $h$.*
- Completeness: *for all $g$, if* $\mathbb{E}[g(Z)|A = a, B] = 0$ *with probability 1, then* $\Pr(g(Z) = 0|A = a) = 1$.

*Let $\mathcal{H}(a)$ be the collection of all $h$ that satisfy* $\mathbb{E}[Y - h(Z)|A = a, B] = 0$ *with probability 1. Then, $\mathcal{H}(a)$ is non-empty, and for all $h \in \mathcal{H}(a)$,* $\mathbb{E}[Y(a)] = \mathbb{E}[h(Z)]$.

Figure 9.6 shows a directed acyclic graph that is consistent with the assumptions of Theorem 9.6. The proxy variables can be seen to be negative control variables in the sense of Shi et al. (2020b), thus making the confounding bridge approach a (double-)negative control approach. Like the primary

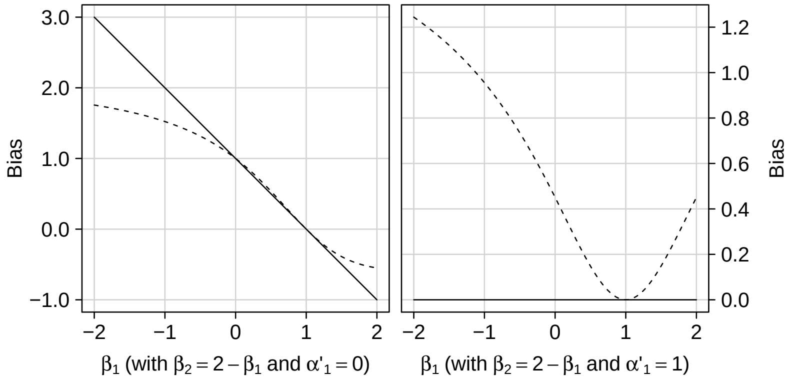


**Figure 9.5:** Illustrating of the potential impact of violating additive or quantile-quantile equi-confounding on the bias of the (generalised) difference-in-difference approach. Solid lines represent the difference-in-difference approach; dashed lines the generalised difference-in-difference.

exposure-outcome association, the exposure-NCO association is confounded by $U$. The function $h$ is referred to as a confounding bridge because the confounding bridge assumption indicates that it links the $Y$-$U$ association with the NCO-$U$ association. The NCE is not part of this link but is meant to help identify it.

The confounding bridge and completeness assumptions can be difficult to grasp. For categorical variables, however, the assumptions are subsumed under the conditions of the next result, due to Miao et al. (2018a) and Shi et al. (2020a).

**Theorem 9.7** (The proximal g-formula for categorical variables)**.** *Let $U, B, Z$ be discrete random variables with finite support such that $U$ has no more categories than $B$ or $Z$. Suppose that for all levels $a$ of $A$, the following conditions hold:*
- Consistency: $Y(a) = Y$ *if* $a = A$.
- Positivity: $0 < \Pr(A = a, B = b)$ *for all categories $b$ of $B$.*
- Latent ignorability: $Y(a) \perp\!\!\!\perp (A, B)|U$ *and* $Z \perp\!\!\!\perp (A, B)|U$.
- Full rank: $\Pr(\boldsymbol{Z}|\boldsymbol{U})$ *and* $\Pr(\boldsymbol{U}|A = a, \boldsymbol{B})$ *have rank equal to the number of levels of $U$.*

*Then,* $\mathbb{E}[Y(a)] = h(\boldsymbol{Z})\Pr(\boldsymbol{Z})$, *where* $h(\boldsymbol{Z}) = \mathbb{E}[Y|A = a, \boldsymbol{B}]\Pr(\boldsymbol{Z}|A = a, \boldsymbol{B})^{-1}$.

Here, following Miao et al. (2018a), for any categorical variables $X, Y, Z$, $\Pr(\boldsymbol{X}|Y, \boldsymbol{Z})$ denotes the matrix of probabilities $\Pr(X = x|Y, \boldsymbol{Z})$ with a one-to-one correspondence between rows and categories $x$ of $X$ and a one-to-one correspondence between columns and categories $z$ of $Z$. Interestingly, the proximal g-formula can also be written as a weighted version of the standard

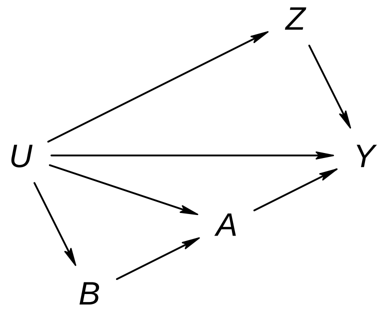


**Figure 9.6:** Causal directed acyclic graph with negative control pair satisfying the latent ignorability condition of Theorem 9.6.

g-formula:

$$\mathbb{E}[Y|A=a,\boldsymbol{B}]\mathrm{diag}(\boldsymbol{W}(a))\Pr(\boldsymbol{B})$$

with weights $\boldsymbol{W}(a) = (\mathrm{diag}\Pr(\boldsymbol{B}))^{-1}\Pr(\boldsymbol{Z}|A=a,\boldsymbol{B})^{-1}\Pr(\boldsymbol{Z})$ and $\mathrm{diag}(\boldsymbol{W}(a))$ and $\mathrm{diag}(\boldsymbol{B})$ denoting the diagonal matrices with main diagonals $\boldsymbol{W}(a)$ and $\boldsymbol{B}$, respectively. In the case that proxy variables $B$ and $Z$ are binary, the expression simplifies to

$$\mathbb{E}\{\mathbb{E}[WY|A=a,B]\}$$

with weights

$$\begin{aligned} W = {} & \frac{(1-B)}{\Pr(B=0)}\frac{\Pr(Z=1|A,B=1)-\Pr(Z=1)}{\Pr(Z=1|A,B=1)-\Pr(Z=1|A,B=0)} \\ & + \frac{-B}{\Pr(B=1)}\frac{\Pr(Z=1|A,B=0)-\Pr(Z=1)}{\Pr(Z=1|A,B=1)-\Pr(Z=1|A,B=0)}. \end{aligned}$$

*Sensitivity to assumption violations*

Theorem 9.7 can accommodate any number of categories of $U$ by taking proxy variables with sufficiently many categories, e.g., by combining sufficiently many proxies. However, upon increasing the number of proxy variables, the latent ignorability assumption becomes more difficult to satisfy in the sense that $Y(a)$ must be independent of increasingly many proxies given $A$ and $U$. In this subsection, we consider the sensitivity of the proximal g-formula for violations of latent ignorability as well as of the assumption that $U$ has no more categories than the proxy variables.

In particular, we consider the case where the variables $A, Y$ of interest and the proxy variables $B, Z$ are binary, where $U$ is a pair $(U_1, U_2)$ of independent binary variables, and where the following models hold:

$$\begin{aligned} U_1 &\sim \mathrm{Bernoulli}(1/2), \\ U_2|U_1 &\sim \mathrm{Bernoulli}(\rho), \\ B|U_1,U_2 &\sim \mathrm{Bernoulli}(\mathrm{expit}\{\alpha_0+U_1+U_2\}), \\ A|U_1,U_2,B &\sim \mathrm{Bernoulli}(\mathrm{expit}\{\beta_0+U_1+\beta_1U_2+B\}), \\ Z|U_1,U_2,B,A &\sim \mathrm{Bernoulli}(\mathrm{expit}\{\gamma_0+U_1-1/2U_2+\gamma_1A\}), \\ Y|U_1,U_2,B,A,Z &\sim \mathrm{Bernoulli}(\mathrm{expit}\{\theta_0+U_1+U_2+Z+\theta_1B\}), \end{aligned}$$

where $\mathrm{expit}(x) = 1/(1+\exp[-x])$ for all $x$. Intercepts $\alpha_0, \beta_0, \gamma_0, \theta_0$ were chosen to ensure that $\Pr(B=1)=\Pr(A=1)=1/2$ and $\Pr(Z=1)=\Pr(Y=1)=1/5$.

We let $\rho = 0, \beta_1 = 1, \gamma_1 = 0, \theta_1 = 0$ by default. In scenario A, instead of taking $\beta_1 = 1, \rho = 0$, we vary $\beta_1$ over $(-4, 4)$ under $\rho = 1/2$ to violate the full rank assumption, which implies that $U$ has no more categories than $B$ or $Z$. In scenario B, instead of taking $\gamma_1 = 0$, we violate the latent ignorability assumption by varying $\gamma_1$ over $(-4, 4)$ (i.e., $Z$ is not a negative control outcome). In scenario C, we violate the same assumption, now by varying $\theta_1$ over $(-4, 4)$ (i.e., $B$ is not a negative control exposure) instead of taking $\theta_1 = 0$.

Figure 9.7 gives the bias of the proximal g-formula for the ATE $\mathbb{E}[Y(1)-Y(0)]$ for all scenarios. Also shown are the differences between the crude risk differences $\mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0]$ and the ATE. The bias is zero under the default parameters, which are consistent with the assumptions of Theorem 9.7. The figure also illustrates that violations of these unverifiable assumptions can have a large impact on the validity of the double-negative control approach.

In an other study, Vlassis et al. (2020) found bias of the crude risk difference to be consistently smaller than that of the proximal g-formula. Our results demonstrate that in some settings, the proximal g-formula results in considerably more bias than what would result from ignoring unmeasured confounding.

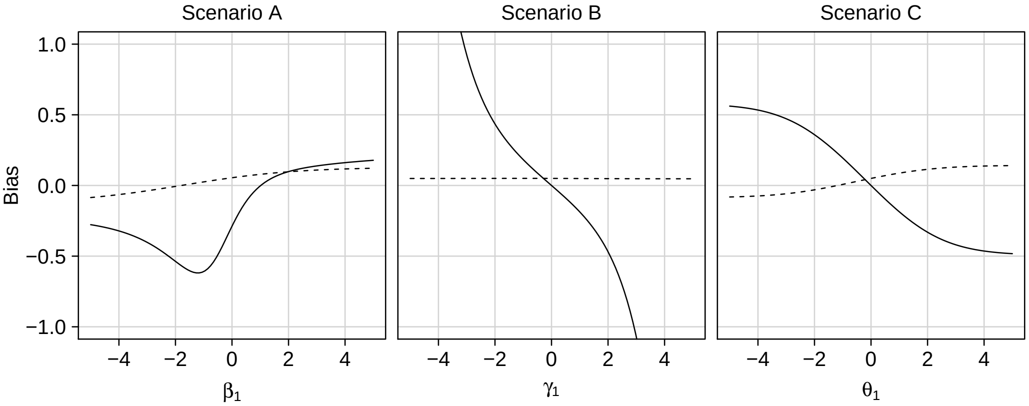


**Figure 9.7:** Bias of crude approach (dashed) and proximal g-formula (solid) under violations of the cardinality assumption (Scenario A), negative control outcome condition (Scenario B), or negative control exposure condition (Scenario C).

## 9.4 Conclusion

Negative controls have gained increasing interest in addressing concerns about the validity of a study. The literature on the topic has tended to consider increasingly ambitious tasks, from confounding detection to full identification of causal effects, typically at the cost of stronger and more complex assumptions. Efforts have been made to introduce negative controls to a broader audience and ensure they are adopted in epidemiological practice (Shi et al., 2020b). However, little attention has yet been given to the methods' assumptions and potential impact of assumptions violations. While the assumptions may be tenable enough in some specific cases to justify an application, in others substantial violations are possible. We have illustrated that assumption violations, some of which are likely even in very simple settings, may have a considerable impact on the validity of the negative control approach, thereby limiting its utility. Despite the possible abundance of negative controls, their routine use in epidemiological practice may fail to strengthen evidence about exposure-outcome effects unless it can be safely assumed that assumption violations are absent or else if the robustness against these violations is well understood. Given the potential impact of assumption violations, it may sometimes be desirable to replace strong conditions for identification with weaker conditions that are easier to verify, even when these weaker conditions imply at most partial identification. Future research in this area may broaden the applicability of negative controls and in turn make them more suited for routine use in epidemiological practice. When they are used, we advise that researches consider the results of their applications carefully and explicitly in light of the methods' limitations and assumptions.

## References

Albert, A. and J. Anderson (1984): "On the existence of maximum likelihood estimates in logistic regression models," *Biometrika*, 71, 1–10.

Arnold, B. F., A. Ercumen, J. Benjamin-Chung, and J. M. Colford Jr (2016): "Brief report: negative controls to detect selection bias and measurement bias in epidemiologic studies," *Epidemiology (Cambridge, Mass.)*, 27, 637.

Birch, M. (1964): "The detection of partial association, I: the 2× 2 case," *Journal of the Royal Statistical Society. Series B (Methodological)*, 313–324.

Flanders, W. D., M. Klein, L. A. Darrow, M. J. Strickland, S. E. Sarnat, J. A. Sarnat, L. A. Waller, A. Winquist, and P. E. Tolbert (2011): "A method for detection of residual confounding in time-series and other observational studies," *Epidemiology (Cambridge, Mass.)*, 22, 59.

Groenwold, R. H. (2013): "Falsification end points for observational studies," *JAMA*, 309, 1769–1771.

Hernán, M. and J. Robins (2020): *Causal Inference: What If*, Boca Raton: Chapman & Hall/CRC.

Lipsitch, M., E. Tchetgen Tchetgen, and T. Cohen (2010): "Negative controls: a tool for detecting confounding and bias in observational studies," *Epidemiology*, 21, 383–388.

Miao, W., Z. Geng, and E. J. Tchetgen Tchetgen (2018a): "Identifying causal effects with proxy variables of an unmeasured confounder," *Biometrika*, 105, 987–993.

Miao, W., X. Shi, and E. Tchetgen Tchetgen (2018b): "A confounding bridge approach for double negative control inference on causal effects," *arXiv e-prints*, arXiv–1808.

Rosenbaum, P. (1989): "The role of known effects in observational studies," *Biometrics*, 45, 557–569.

Shi, X., W. Miao, J. C. Nelson, and E. J. Tchetgen Tchetgen (2020a): "Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 521–540.

Shi, X., W. Miao, and E. Tchetgen Tchetgen (2020b): "A selective review of negative control methods in epidemiology," *Current Epidemiology Reports*, 1–13.

Sofer, T., D. B. Richardson, E. Colicino, J. Schwartz, and E. J. Tchetgen Tchetgen (2016): "On negative outcome control of unobserved confounding as a generalization of difference-in-differences," *Statistical science: a review journal of the Institute of Mathematical Statistics*, 31, 348.

Tchetgen, E. J. T., A. Ying, Y. Cui, X. Shi, and W. Miao (2020): "An introduction to proximal causal learning," *arXiv preprint arXiv:2009.10982*.

Tchetgen Tchetgen, E. (2013): “The control outcome calibration approach for causal inference with unobserved confounding,” *American Journal of Epidemiology*, 179, 633–640.

Vlassis, N., P. Hebda, S. McBride, and A. Noulas (2020): “On proximal causal learning with many hidden confounders,” *arXiv preprint arXiv:2012.06725*.

## Supplementary Material

### S9.1 Identifiability of the direction of bias using an NCO/NCE

**Theorem** (Identification of the direction of bias using an NCO/NCE)**.** *Suppose the following conditions hold:*

- Latent ignorability for some scalar $U$: $Z \perp\!\!\!\perp (A, Y)|U$ *and* $Y \perp\!\!\!\perp A|U$.
- Primary exposure model: $A = \alpha_0 + \alpha_1 U + \epsilon$, $\epsilon \perp\!\!\!\perp U$, $\mathbb{E}[\epsilon] = 0$.
- Primary outcome model: $Y = \gamma_0 + \gamma_1 U + \theta A + \varepsilon$, $\varepsilon \perp\!\!\!\perp (A, U)$, $\mathbb{E}[\varepsilon] = 0$.
- NCO/NCE model: $Z = \beta_0^* + \beta_1^* U + \delta$, $\delta \perp\!\!\!\perp (A, Y, U), \mathbb{E}[\delta] = 0$.

*Then,* $\hat{\theta} - \theta$ *has the same sign as*

$$\frac{\text{Cov}(Y, Z)}{\text{Cov}(A, Z)} - \hat{\theta}.$$

*Proof.* For the ordinary least squares coefficient $\hat{\theta} = \text{Cov}(Y, A)/\text{Var}(A)$ in the regression of $Y$ on $A$, we have

$$\begin{aligned} \hat{\theta} - \theta &= \gamma_1 \frac{\text{Cov}(U, A)}{\text{Var}(A)} && \text{(by the primary outcome model)} \\ &= \gamma_1 \alpha_1 \frac{\text{Var}(U)}{\text{Var}(A)}. && \text{(by the primary exposure model)} \end{aligned}$$

Note that $\text{Var}(\mathbb{E}[A|U]) = \alpha_1^2 \text{Var}(U)$ and, by the law of total variance, $\text{Var}(A) = \text{Var}(\mathbb{E}[A|U]) + \mathbb{E}[\text{Var}(A|U)]$. Thus, $\text{Var}(A) - \mathbb{E}[\text{Var}(A|U)] = \alpha_1^2 \text{Var}(U)$

$$\hat{\theta} - \theta = \frac{\gamma_1}{\alpha_1} \frac{\text{Var}(A) - \mathbb{E}[\text{Var}(A|U)]}{\text{Var}(A)}.$$

The fraction $(\text{Var}(A) - \mathbb{E}[\text{Var}(A|U)])/\text{Var}(A)$ can be interpreted as the proportion of variance of $A$ that is explained by $U$. By the law of total variance, the fraction is bounded by 0 and 1.

Next, note that

$$\begin{aligned} \frac{\text{Cov}(Y, Z)}{\text{Cov}(A, Z)} &= \frac{\gamma_1 \text{Cov}(Z, U) + \theta \text{Cov}(A, Z)}{\text{Cov}(A, Z)} && \text{(by the primary outcome model)} \\ &= \gamma_1 \frac{\text{Cov}(Z, U)}{\text{Cov}(A, Z)} + \theta \\ &= \gamma_1 \frac{\text{Cov}(Z, U)}{\alpha_1 \text{Cov}(U, Z)} + \theta && \text{(by the primary exposure model)} \end{aligned}$$

$$= \frac{\gamma_1}{\alpha_1} + \theta.$$

Hence,

$$\begin{aligned}
\hat{\theta} - \theta &= \left(\frac{\text{Cov}(Y,Z)}{\text{Cov}(A,Z)} - \theta\right)\frac{\text{Var}(A) - \mathbb{E}[\text{Var}(A|U)]}{\text{Var}(A)}, \\
\theta &= \frac{\hat{\theta} - \lambda\dfrac{\text{Cov}(Y,Z)}{\text{Cov}(A,Z)}}{1-\lambda}, \\
\hat{\theta} - \theta &= \left(\frac{\text{Cov}(Y,Z)}{\text{Cov}(A,Z)} - \hat{\theta}\right)\frac{\lambda}{1-\lambda},
\end{aligned}$$

where $\lambda = (\text{Var}(A) - \mathbb{E}[\text{Var}(A|U)])/\text{Var}(A)$. Clearly, since $\lambda \in [0,1]$, the sign of the bias $\hat{\theta} - \theta$ is identified by $\text{Cov}(Y,Z)/\text{Cov}(A,Z) - \hat{\theta}$. □

**No identifiability of the direction of bias when $Z$ is not an NCE.** *Consider the following models*

$$\begin{aligned}
&U \sim \text{Normal}(\mathbb{E}[U], \text{Var}(U)), \\
&A = \alpha_0 + \alpha_1 U + \epsilon,\ \epsilon|U \sim \text{Normal}(0, \text{Var}(\epsilon)), \\
&Z = \beta_0^* + \beta_1^* U + \delta,\ \delta|(U,A) \sim \text{Normal}(0, \text{Var}(\delta)), \\
&Y = \gamma_0 + \gamma_1 U + \gamma_2 Z + \theta A + \varepsilon,\ \varepsilon|(Z,A,U) \sim \text{Normal}(0, \text{Var}(\varepsilon)),
\end{aligned}$$

*which are compatible with those of the above theorem if $\gamma_2 = 0$. If $\gamma_2 \neq 0$, then the Latent ignorability condition is violated because $Z \not\perp\!\!\!\perp (A,Y)|U$. If it were possible to infer from the distribution of the observables the direction of bias $\hat{\theta} - \theta$, then there exists some function $g$ of the joint distribution $F$ of $(A,Y,Z)$ such that $g(F)[\hat{\theta} - \theta] > 0$. To prove that this is false, it suffices to show that for some $F$, the bias $\hat{\theta} - \theta$ may be positive and negative, depending on unobservables, so that for all $g$, we have $g(F)[\hat{\theta} - \theta] \ngtr 0$.*

*Consider the models of the previous section with parameters set to the following values to yield multivariate normal distributions $G, H$:*

|  | $G$ | $H$ |  | $G$ | $H$ |
|---|---|---|---|---|---|
| $\mathrm{Var}(U)$ | 1 | 1 | $\alpha_1$ | 1 | 1 |
| $\mathbb{E}[U]$ | 0 | 0 | $\beta_1^*$ | 1 | 1 |
| $\mathrm{Var}(\epsilon)$ | 1 | 1 | $\gamma_1$ | $-5$ | 1 |
| $\alpha_0$ | 0 | 0 | $\theta$ | 1 | $-1$ |
| $\mathrm{Var}(\delta)$ | 1 | 1 | $\gamma_2$ | 3 | 1 |
| $\beta_0^*$ | 0 | 0 |  |  |  |
| $\mathrm{Var}(\varepsilon)$ | 1 | 9 |  |  |  |
| $\gamma_0$ | 0 | 0 |  |  |  |

*Given zero means of $A, Y, Z$, the corresponding covariance matrices*

$$\mathrm{Cov}(G) = \begin{bmatrix} 1 & 1 & 1 & -1 \\ 1 & 2 & 1 & 0 \\ 1 & 1 & 2 & 2 \\ -1 & 0 & 2 & 12 \end{bmatrix}, \qquad \mathrm{Cov}(H) = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 0 \\ 1 & 1 & 2 & 2 \\ 1 & 0 & 2 & 12 \end{bmatrix}$$

*imply the same distribution for $(A, Y, Z)$, despite the fact that the true effects $\theta$ have opposite signs. This shows that in general the direction of bias cannot be identified.*

## S9.2 Proofs to theorems in section 9.3

*Proof to Theorem 9.1.* For all $a$,

$$\begin{aligned} g^{-1}(\beta_0 + \beta_2 y) &= \mathbb{E}[Z|A = 0, Y = y] && \text{(by NCO model)} \\ &= \mathbb{E}[Z|A = 0, Y(0) = y] && \text{(by consistency)} \\ &= \mathbb{E}[Z|A = a, Y(0) = y] \\ & \text{(by exposure-NCO independence given counterfactual outcome)} \\ &= \mathbb{E}[Z|A = a, Y(0) + \theta A - \theta A = y] \\ &= \mathbb{E}[Z|A = a, Y(a) = y + \theta a] && \text{(by rank preservation)} \\ &= \mathbb{E}[Z|A = a, Y = y + \theta a] && \text{(by consistency)} \\ &= g^{-1}(\beta_0 + \beta_1 a + \beta_2(y + \theta a)) \\ &= g^{-1}(\beta_0 + (\beta_1 + \beta_2\theta)a + \beta_2 y), \end{aligned}$$

so that, for $a = 1$,

$$\beta_0 + \beta_2 y = \beta_0 + (\beta_1 + \beta_2\theta) + \beta_2 y,$$

$$\theta = -\beta_1/\beta_2.$$

□

*Proof to Theorem 9.2.*

$$\begin{aligned}
Z &= \beta_0 + \beta_1 Y(0) + \rho(A - \mathbb{E}[A|Y(0)]) + \chi \quad \text{(by linear NCO model)}\\
&= \beta_0 + \beta_1 Y(0) + \rho(A - \alpha_0 - \alpha_1 Y(0)) + \chi\\
&\qquad\qquad \text{(by linear exposure model)}\\
&= \beta_0 + \beta_1(Y(A) - \theta A) + \rho(A - \alpha_0 - \alpha_1 Y(A) + \alpha_1 \theta A) + \chi\\
&\qquad\qquad \text{(by rank preservation)}\\
&= (\beta_0 - \rho\alpha_0) + (\rho + [\rho\alpha_1 - \beta_1]\theta)A + (\beta_1 - \rho\alpha_1)Y + \chi,\\
&\qquad\qquad \text{(by consistency)}\\
\text{and } \mathbb{E}[Z|A,Y] &= (\beta_0 - \rho\alpha_0 + \mathbb{E}[\chi]) + (\rho + [\rho\alpha_1 - \beta_1]\theta)A + (\beta_1 - \rho\alpha_1)Y,\\
&\qquad\qquad \text{(by linear NCO model)}
\end{aligned}$$

so that

$$\beta_1^* = \rho + (\rho\alpha_1 - \beta_1)\theta \text{ and } \beta_2^* = \beta_1 - \rho\alpha_1,$$

and, in turn, $\theta = (\beta_1^* - \rho)/\beta_2^*$. □

*Proof to Theorem 9.3.* We have

$$\begin{aligned}
&\mathbb{E}[Z|A = 1-a]\\
&\quad= \mathbb{E}\{\mathbb{E}[Z|A = 1-a, Y(a)]|A = 1-a\}\\
&\quad= \mathbb{E}\{\mathbb{E}[Z|A = a, Y(a)]|A = 1-a\}\\
&\qquad\quad \text{(by exposure-NCO independence given counterfactual outcome)}\\
&\quad= \mathbb{E}[Z|A = a, Y = 0]\Pr(Y(a) = 0|A = 1-a) +\\
&\qquad + \mathbb{E}[Z|A = a, Y = 1]\Pr(Y(a) = 1|A = 1-a) \qquad \text{(by consistency)}\\
&\quad= \mathbb{E}[Z|A = a, Y = 0] + \{\mathbb{E}[Z|A = a, Y = 1] - \mathbb{E}[Z|A = a, Y = 0]\}\\
&\qquad \times \Pr(Y(a) = 1|A = 1-a),
\end{aligned}$$

so that

$$\Pr(Y(a) = 1|A = 1-a) = \frac{\mathbb{E}[Z|A = 1-a] - \mathbb{E}[Z|A = a, Y = 0]}{\mathbb{E}[Z|A = a, Y = 1] - \mathbb{E}[Z|A = a, Y = 0]}.$$

It follows that

$$\mathbb{E}[Y(a)] = \mathbb{E}[Y(a)|A = a]\Pr(A = a) + \mathbb{E}[Y(a)|A = 1-a]\Pr(A = 1-a)$$

$$
\begin{aligned}
&= \mathbb{E}[Y|A=a]\Pr(A=a) + \mathbb{E}[Y(a)|A=1-a]\Pr(A=1-a) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by consistency)} \\
&= \mathbb{E}[Y|A=a]\Pr(A=a) \\
&\quad + \frac{\mathbb{E}[Z|A=1-a] - \mathbb{E}[Z|A=a,Y=0]}{\mathbb{E}[Z|A=a,Y=1] - \mathbb{E}[Z|A=a,Y=0]}\Pr(A=1-a).
\end{aligned}
$$

□

*Proof to Theorem 9.4.*

$$
\begin{aligned}
&\mathbb{E}[Y(1)-Y(0)|A=1] \\
&\quad= \mathbb{E}[Y(1)|A=1] - \mathbb{E}[Y(0)|A=1] \\
&\quad= \mathbb{E}[Y|A=1] - \mathbb{E}[Y(0)|A=1] \qquad \text{(by consistency)} \\
&\quad= \mathbb{E}[Y|A=1] - (\mathbb{E}[Y(0)|A=1] - \mathbb{E}[Y(0)|A=0]) - \mathbb{E}[Y(0)|A=0] \\
&\quad= \mathbb{E}[Y|A=1] - (\mathbb{E}[N|A=1] - \mathbb{E}[N|A=0]) - \mathbb{E}[Y(0)|A=0] \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(by additive equi-confounding)} \\
&\quad= (\mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0]) - (\mathbb{E}[N|A=1] - \mathbb{E}[N|A=0]) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(by consistency)}
\end{aligned}
$$

□

*Proof to Theorem 9.5.* By quantile-quantile equi-confounding, we have, for all $p \in [0,1]$,

$$
\begin{aligned}
F_0(F_1^{-1}(p)) &= G_0(G_1^{-1}(p)), \\
F_0^{-1}(F_0(F_1^{-1}(p))) &= F_0^{-1}(G_0(G_1^{-1}(p))), \\
F_1^{-1}(p) &= F_0^{-1}(G_0(G_1^{-1}(p))). \qquad \text{(under strictly monotonic } F_1\text{)}
\end{aligned}
$$

Note that the right-hand side of the above equality is a functional of observables because $F_0(y) = \Pr(Y(0) \le y|A=0) = \Pr(Y \le y|A=0)$ by consistency. Now, letting $V \sim \text{Uniform}[0,1]$, we have that $F_1^{-1}(V) \sim Y(0)|A=1$ by the Probability Integral Transform theorem, and so

$$
\mathbb{E}[Y(0)|A=1] = \mathbb{E}[F_0^{-1}(G_0(G_1^{-1}(V)))].
$$

□

*Proof to Theorem 9.6.* Let $h$ be the function satisfying $\mathbb{E}[Y|A=a,U] = \mathbb{E}[h(Z)|A=a,U]$ with probability 1 (and note that this function exists by the confounding bridge assumption). Let $\mathcal{U} = \{u : \mathbb{E}[Y|A = a, U = u] = \mathbb{E}[h(Z)|A = a, U = u]\}$, so that $\Pr(U \in \mathcal{U}) = 1$ and

$$\begin{aligned}
\mathbb{E}[Y(a)] &= \mathbb{E}[Y(a)|U \in \mathcal{U}] \\
&= \mathbb{E}\{\mathbb{E}[Y(a)|U]|U \in \mathcal{U}\} \\
&= \mathbb{E}\{\mathbb{E}[Y(a)|A = a, U]|U \in \mathcal{U}\} \\
&\qquad\qquad \text{(since } Y(a) \perp\!\!\!\perp A|U \text{ by latent ignorability)} \\
&= \mathbb{E}\{\mathbb{E}[Y|A = a, U]|U \in \mathcal{U}\} \qquad \text{(by consistency)} \\
&= \mathbb{E}\{\mathbb{E}[h(Z)|A = a, U]|U \in \mathcal{U}\} \\
&= \mathbb{E}\{\mathbb{E}[h(Z)|U]|U \in \mathcal{U}\} \qquad \text{(since } Z \perp\!\!\!\perp A|U \text{ by latent ignorability)} \\
&= \mathbb{E}[h(Z)|U \in \mathcal{U}] \\
&= \mathbb{E}[h(Z)].
\end{aligned}$$

Next, note that by the confounding bridge assumption, for all $U \in \mathcal{U}$,

$$\begin{aligned}
\mathbb{E}[Y|A = a, U] &= \mathbb{E}[h(Z)|A = a, U] \\
\mathbb{E}[Y|A = a, B, U] &= \mathbb{E}[h(Z)|A = a, B, U], \qquad \text{(by latent ignorability)}
\end{aligned}$$

so that

$$\begin{aligned}
\mathbb{E}\{\mathbb{E}[Y|A = a, B, U]|A = a, B\} &= \mathbb{E}\{\mathbb{E}[h(Z)|A = a, B, U]|A = a, B\} \\
\mathbb{E}[Y|A = a, B] &= \mathbb{E}[h(Z)|A = a, B], \\
\mathbb{E}[Y - h(Z)|A = a, B] &= 0.
\end{aligned}$$

Let $\mathcal{H}(a)$ be the collection of all $h'$ satisfying $\mathbb{E}[Y - h'(Z)|A = a, B] = 0$ with probability 1. Now, for any $h' \in \mathcal{H}(a)$, we must have

$$\mathbb{E}[h(Z) - h'(Z)|A = a, B] = 0.$$

But from completeness, with $g(Z) = h(Z) - h'(Z)$, it follows that $h(Z) = h'(Z)$ with probability 1. This concludes the proof. □

*Proof to Theorem 9.7.* Since $Z \perp\!\!\!\perp (A, B)|U$, we have $\Pr(\boldsymbol{Z}|A = a, \boldsymbol{B}) = \Pr(\boldsymbol{Z}|\boldsymbol{U}) \Pr(\boldsymbol{U}|A = a, \boldsymbol{B})$. Since matrices $\Pr(\boldsymbol{Z}|\boldsymbol{U})$ and $\Pr(\boldsymbol{U}|A = a, \boldsymbol{B})$ are of full rank, $\Pr(\boldsymbol{Z}|A = a, \boldsymbol{B})$ is of full rank and has left or right inverse $\Pr(\boldsymbol{Z}|A = a, \boldsymbol{B})^{-1}$. Let $h(\boldsymbol{Z}) = \mathbb{E}[Y|A = a, \boldsymbol{B}] \Pr(\boldsymbol{Z}|A = a, \boldsymbol{B})^{-1}$ and observe that

$$h(\boldsymbol{Z}) = \mathbb{E}[Y(a)|A = a, \boldsymbol{B}] \Pr(\boldsymbol{Z}|A = a, \boldsymbol{B})^{-1} \qquad \text{(by consistency)}$$

$$
\begin{aligned}
&= \mathbb{E}[Y(a)|\boldsymbol{U}] \Pr(\boldsymbol{U}|A = a, \boldsymbol{B}) \Pr(\boldsymbol{Z}|A = a, \boldsymbol{B})^{-1} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(since } Y(a) \perp\!\!\!\perp (A, B)|U) \\
&= \mathbb{E}[Y(a)|\boldsymbol{U}] \Pr(\boldsymbol{U}|A = a, \boldsymbol{B})[\Pr(\boldsymbol{Z}|\boldsymbol{U}) \Pr(\boldsymbol{U}|A = a, \boldsymbol{B})]^{-1} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(since } Z \perp\!\!\!\perp (A, B)|U) \\
&= \mathbb{E}[Y(a)|\boldsymbol{U}] \Pr(\boldsymbol{U}|A = a, \boldsymbol{B}) \Pr(\boldsymbol{U}|A = a, \boldsymbol{B})^{-1} \Pr(\boldsymbol{Z}|\boldsymbol{U})^{-1} \\
&= \mathbb{E}[Y(a)|\boldsymbol{U}] \Pr(\boldsymbol{Z}|\boldsymbol{U})^{-1}.
\end{aligned}
$$

It follows that $\mathbb{E}[Y(a)|\boldsymbol{U}] = h(\boldsymbol{Z}) \Pr(\boldsymbol{Z}|\boldsymbol{U})$ and in turn $\mathbb{E}[Y(a)] = \mathbb{E}[Y(a)|\boldsymbol{U}] \Pr(\boldsymbol{U}) = h(\boldsymbol{Z}) \Pr(\boldsymbol{Z})$, as desired.

□

## S9.3 Derivation of expressions in section 9.3.1

### *S9.3.1 Implications of models* (9.1)

*Expression of the COCA*

An implementation of the COCA by ordinary least squares under the linear NCO model $\mathbb{E}[Z|A, Y] = \beta_0 + \beta_1 A + \beta_2 Y$, identifies the following quantity

$$
\begin{aligned}
\hat{\theta} &= -\frac{\hat{\beta}_1}{\hat{\beta}_2} \\
&= -\frac{\mathrm{Cov}(A, Z)\mathrm{Var}(Y) - \mathrm{Cov}(Y, Z)\mathrm{Cov}(A, Y)}{\mathrm{Cov}(Y, Z)\mathrm{Var}(A) - \mathrm{Cov}(A, Z)\mathrm{Cov}(A, Y)},
\end{aligned}
$$

where

$$
\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{Var}(A)\alpha_1^2 + \sigma_Y^2 + \sigma_\theta^2 \mathrm{Var}(A) + \sigma_\theta^2 \mathbb{E}[A]^2 + \mathbb{E}[\theta]^2 \mathrm{Var}(A) \\
&\quad + 2\mathrm{Var}(A)\alpha_1 \mathbb{E}[\theta], \\
\mathrm{Var}(Z) &= \mathrm{Var}(A)\alpha_1^2\gamma_1^2 + \gamma_1^2\sigma_Y^2 + \sigma_Z^2, \\
\mathrm{Cov}(A, Y) &= \mathrm{Var}(A)\alpha_1 + \mathrm{Var}(A)\mathbb{E}[\theta], \\
\mathrm{Cov}(A, Z) &= \mathrm{Var}(A)\alpha_1\gamma_1, \\
\mathrm{Cov}(Y, Z) &= \gamma_1(\mathrm{Var}(A)\alpha_1^2 + \sigma_Y^2 + \mathrm{Var}(A)\alpha_1 \mathbb{E}[\theta]).
\end{aligned}
$$

*Deterministic relation between $\mathbb{E}[\theta]$ and $\mathrm{Var}[\theta]$ given observed data distribution*

From the expressions of the variances and covariates above, for arbitrary $\mathbb{E}[\theta], \mathrm{Var}(A)$, it follows that

$$\begin{aligned}
\alpha_1 &= \frac{\mathrm{Cov}(A,Y) - \mathrm{Var}(A)\mathbb{E}[\theta]}{\mathrm{Var}(A)},\\
\alpha_0 &= \mathbb{E}[Y] - (\alpha_1 + \mathbb{E}[\theta])\mathbb{E}[A],\\
\gamma_1 &= \frac{\mathrm{Cov}(A,Z)}{\mathrm{Var}(A)\alpha_1},\\
\gamma_0 &= \mathbb{E}[Z] - (\alpha_0\gamma_1 + \alpha_1\gamma_1\mathbb{E}[A]),\\
\sigma_Y^2 &= \frac{\mathrm{Cov}(Y,Z) - \gamma_1(\mathrm{Var}(A)\alpha_1^2 + \mathrm{Var}(A)\alpha_1\mathbb{E}[\theta])}{\gamma_1},\\
\sigma_\theta^2 &= \frac{\mathrm{Var}(Y) - [\mathrm{Var}(A)\alpha_1^2 + \sigma_Y^2 + \mathrm{Var}(A)\mathbb{E}[\theta]^2 + 2\mathrm{Var}(A)\alpha_1\mathbb{E}[\theta]]}{\mathrm{Var}(A) + \mathbb{E}[A]^2},\\
\sigma_Z^2 &= \mathrm{Var}(Z) - (\mathrm{Var}(A)\alpha_1^2\gamma_1^2 + \gamma_1^2\sigma_Y^2),
\end{aligned}$$

provided that $\mathrm{Var}(A), \alpha_1, \gamma_1 \neq 0$, and $\sigma_Y^2, \sigma_\theta^2, \sigma_Z^2 \geq 0$. Note that the right-hand sides of every equality are expressed only in terms of functionals of the available data distribution and the left-hand sides of the equalities above it. It follows that we have a deterministic relationship between $\mathrm{Var}(\theta)$ and $\mathbb{E}[\theta]$ given the observed data distribution of $(A, Y, Z)$. In fact, the relationship is linear:

$$\begin{aligned}
\sigma_\theta^2 &= \frac{\mathrm{Var}(Y)\mathrm{Cov}(A,Z) - \mathrm{Cov}(A,Y)\mathrm{Cov}(Y,Z)}{(\mathrm{Var}(A) + \mathbb{E}[A]^2)\mathrm{Cov}(A,Z)}\\
&\quad - \frac{\mathrm{Cov}(A,Y)\mathrm{Cov}(A,Z) - \mathrm{Var}(A)\mathrm{Cov}(Y,Z)}{(\mathrm{Var}(A) + \mathbb{E}[A]^2)\mathrm{Cov}(A,Z)}\mathbb{E}[\theta].\\
&= \frac{\mathrm{Var}(A)\mathrm{Var}(Y) - \mathrm{Cov}(A,Y)^2}{(\mathrm{Var}(A) + \mathbb{E}[A]^2)\mathrm{Cov}(A,Z)}(\hat{\beta}_1 - \hat{\beta}_2\mathbb{E}[\theta]).
\end{aligned}$$

*The distribution of $Z|A, Y$*

First note that

$$\begin{aligned}
\mathbb{E}[\theta|Y, A=0] &= \mathbb{E}[\theta|Y(0), A=0] && \text{(by consistency)}\\
&= \mathbb{E}[\theta|A=0] && \text{(since } Y(0) \perp\!\!\!\perp \theta|A)\\
&= \mathbb{E}[\theta]. && \text{(since } \theta \perp\!\!\!\perp A)
\end{aligned}$$

Next, for arbitrary $a$, consider $\mathbb{E}[\theta|Y, A = a]$ and note that $(\theta, Y)|A = a$ takes a bivariate normal distribution with means

$$\begin{aligned}
\mathbb{E}[\theta|A = a] &= \mathbb{E}[\theta], \\
\mathbb{E}[Y|A = a] &= \mathbb{E}[Y(A)|A = a] && \text{(by consistency)} \\
&= \mathbb{E}[Y(0) + \theta A|A = a] \\
&= \alpha_0 + \alpha_1 a + \mathbb{E}[\theta],
\end{aligned}$$

variances

$$\begin{aligned}
\mathrm{Var}(\theta|A = a) &= \sigma_\theta^2, && \text{(since } \theta \perp\!\!\!\perp A\text{)} \\
\mathrm{Var}(Y|A = a) &= \mathrm{Var}(Y(A)|A = a) && \text{(by consistency)} \\
&= \mathrm{Var}(Y(0) + \theta A|A = a) \\
&= \mathrm{Var}(Y(0)|A = a) + \mathrm{Var}(\theta) && \text{(since } Y(0) \perp\!\!\!\perp \theta|A \text{ and } \theta \perp\!\!\!\perp A\text{)} \\
&= \sigma_Y^2 + \sigma_\theta^2,
\end{aligned}$$

and correlation

$$\begin{aligned}
&\mathrm{Cor}(Y, \theta|A = a) \\
&\quad = \sqrt{\frac{\mathrm{Cov}^2(Y, \theta|A = a)}{\mathrm{Var}(Y|A = a)\mathrm{Var}(\theta|A = a)}} \\
&\quad = \sqrt{\frac{\mathbb{E}[(Y - \mathbb{E}[Y|A = a])(\theta - \mathbb{E}[\theta|A = a])|A = a]^2}{\mathrm{Var}(Y|A = a)\mathrm{Var}(\theta|A = a)}} \\
&\quad = \sqrt{\frac{\mathbb{E}[(Y(A) - \mathbb{E}[Y(A)|A = a])(\theta - \mathbb{E}[\theta|A = a])|A = a]^2}{\mathrm{Var}(Y|A = a)\mathrm{Var}(\theta|A = a)}} && \text{(by consistency)} \\
&\quad = \sqrt{\frac{\mathbb{E}[(Y(0) + \theta A - \mathbb{E}[Y(0) + \theta A|A = a])(\theta - \mathbb{E}[\theta|A = a])|A = a]^2}{\mathrm{Var}(Y|A = a)\mathrm{Var}(\theta|A = a)}} \\
&\quad = \sqrt{\frac{\begin{array}{l}\mathbb{E}[((Y(0) - \mathbb{E}[Y(0)|A = a]) \\ \quad + (\theta A - \mathbb{E}[\theta|A = a]))(\theta - \mathbb{E}[\theta|A = a])|A = a]^2\end{array}}{\mathrm{Var}(Y|A = a)\mathrm{Var}(\theta|A = a)}} \\
&\quad = \sqrt{\frac{[\mathrm{Cov}(Y(0), \theta|A = a) + \mathrm{Cov}(\theta A, \theta|A = a)]^2}{\mathrm{Var}(Y|A = a)\mathrm{Var}(\theta|A = a)}} \\
&\quad = \sqrt{\frac{a^2\mathrm{Var}(\theta|A = a)^2}{\mathrm{Var}(Y|A = a)\mathrm{Var}(\theta|A = a)}} && \text{(since } Y(0) \perp\!\!\!\perp \theta|A\text{)}
\end{aligned}$$

$$= a\sqrt{\frac{\sigma_\theta^2}{\sigma_Y^2 + \sigma_\theta^2}}.$$

Therefore,

$$\mathbb{E}[\theta|Y, A = a] = \mathbb{E}[\theta] + \sqrt{\frac{\sigma_\theta^2}{\sigma_Y^2 + \sigma_\theta^2}}\frac{\sigma_\theta^2}{\sigma_Y^2}[-(\alpha_0 + \mathbb{E}[\theta])a - \alpha_1 a^2 + aY]$$

(DeGroot and Schervisch, 2012, Theorem 5.10.4, p. 340).

Hence,

$$\mathbb{E}[\theta|Y, A] = \mathbb{E}[\theta] + \sqrt{\frac{\sigma_\theta^2}{\sigma_Y^2 + \sigma_\theta^2}}\frac{\sigma_\theta^2}{\sigma_Y^2}[-(\alpha_0 + \mathbb{E}[\theta])A - \alpha_1 A^2 + AY]$$

and

$$\begin{aligned} Z &= \gamma_0 + \gamma_1 Y(0) + \varepsilon,\ \varepsilon|(A, \theta, Y(0)) \sim \text{Normal}(0, \sigma_Z^2) \\ &= \gamma_0 + \gamma_1(Y(A) - \theta A) + \varepsilon \\ &= \gamma_0 + \gamma_1(Y - \theta A) + \varepsilon && \text{(by consistency)} \\ &= \gamma_0 - \gamma_1\theta A + \gamma_1 Y + \varepsilon, \end{aligned}$$

so that $Z|A, Y$ has a normal distribution with mean

$$\begin{aligned} \mathbb{E}[Z|A, Y] &= \gamma_0 - \gamma_1\mathbb{E}[\theta|Y, A]A + \gamma_1 Y \\ &= \gamma_0 - \gamma_1 A\left[\mathbb{E}[\theta] + \sqrt{\frac{\sigma_\theta^2}{\sigma_Y^2 + \sigma_\theta^2}}\frac{\sigma_\theta^2}{\sigma_Y^2}[-(\alpha_0 + \mathbb{E}[\theta])A - \alpha_1 A^2 + AY]\right] \\ &\quad + \gamma_1 Y \\ &= \gamma_0 - \gamma_1\mathbb{E}[\theta]A + \sqrt{\frac{\sigma_\theta^2}{\sigma_Y^2 + \sigma_\theta^2}}\frac{\sigma_\theta^2}{\sigma_Y^2}[(\alpha_0 + \mathbb{E}[\theta])\gamma_1 A^2 \\ &\quad + \alpha_1\gamma_1 A^3 - \gamma_1 A^2 Y] + \gamma_1 Y \\ &= \beta_0^* + \beta_1^* A + \beta_2^* A^2 + \beta_3^* A^3 + \beta_4^* Y + \beta_5^* A^2 Y, \end{aligned}$$

where

$$\begin{aligned} \beta_0^* &= \gamma_0, \\ \beta_1^* &= -\gamma_1\mathbb{E}[\theta], \\ \beta_2^* &= \sqrt{\frac{\sigma_\theta^2}{\sigma_Y^2 + \sigma_\theta^2}}\frac{\sigma_\theta^2}{\sigma_Y^2}(\alpha_0 + \mathbb{E}[\theta])\gamma_1, \end{aligned}$$

$$\begin{aligned}
\beta_3^* &= \sqrt{\frac{\sigma_\theta^2}{\sigma_Y^2 + \sigma_\theta^2}\frac{\sigma_\theta^2}{\sigma_Y^2}}\alpha_1\gamma_1, \\
\beta_4^* &= \gamma_1, \\
\beta_5^* &= -\sqrt{\frac{\sigma_\theta^2}{\sigma_Y^2 + \sigma_\theta^2}\frac{\sigma_\theta^2}{\sigma_Y^2}}\gamma_1,
\end{aligned}$$

so $\mathbb{E}[\theta] = -\beta_1^*/\beta_4^*$ if $\beta_4^* \neq 0$. Therefore, with a continuous primary outcome and non-binary exposure, the rank preservation assumption can sometimes be dropped whilst maintaining identifiability. If $A$ is binary, we have $\mathbb{E}[Z|A, Y] = \beta_0^* + (\beta_1^* + \beta_2^* + \beta_3^*)A + \beta_4^* Y + \beta_5^* AY$, where

$$\begin{aligned}
(\beta_1^* + \beta_2^* + \beta_3^*) &= \gamma_1\sqrt{\frac{\sigma_\theta^2}{\sigma_Y^2 + \sigma_\theta^2}\frac{\sigma_\theta^2}{\sigma_Y^2}}(\alpha_0 + \alpha_1) + \gamma_1\left(\sqrt{\frac{\sigma_\theta^2}{\sigma_Y^2 + \sigma_\theta^2}\frac{\sigma_\theta^2}{\sigma_Y^2}} - 1\right)\mathbb{E}[\theta] \\
&= -\beta_5^*(\alpha_0 + \alpha_1) - (\beta_4^* + \beta_5^*)\mathbb{E}[\theta].
\end{aligned}$$

This suggests a test for violations of rank preservation since the interaction term coefficient $\beta_5^*$ is zero if and only if $\text{Var}(\theta) = 0$ or $\beta_4^* = 0$. Provided that $\beta_4^* \neq 0$, a valid test of the null hypothesis $\beta_5^* = 0$ is thus a valid test of rank preservation under the above models.

### *S9.3.2 Implications of models* (9.3)

Under models 9.3, we have the following variances and covariances:

$$\begin{aligned}
\text{Var}(A) &= \alpha_1^2\text{Var}(U_1) + \alpha_2^2\text{Var}(U_2) + \sigma_A^2, \\
\text{Var}(Y) &= (1 + \theta\alpha_1)^2\text{Var}(U_1) + (1 + \theta\alpha_2)^2\text{Var}(U_2) + \theta^2\sigma_A^2, \\
\text{Var}(Z) &= (\alpha_1')^2\text{Var}(U_1) + (\alpha_2')^2\text{Var}(U_1) + \sigma_Z^2, \\
\text{Cov}(A, Y) &= (1 + \theta\alpha_1)\alpha_1\text{Var}(U_1) + (1 + \theta\alpha_2)\alpha_2\text{Var}(U_2) + \theta\sigma_A^2, \\
\text{Cov}(A, Z) &= \alpha_1\alpha_1'\text{Var}(U_1) + \alpha_2\alpha_2'\text{Var}(U_2), \\
\text{Cov}(Y, Z) &= (1 + \theta\alpha_1)\alpha_1'\text{Var}(U_1) + (1 + \theta\alpha_2)\alpha_2'\text{Var}(U_2)
\end{aligned}$$

and means

$$\begin{aligned}
\mathbb{E}[A] &= \alpha_0 + \alpha_1\mathbb{E}[U_1] + \alpha_2\mathbb{E}[U_2], \\
\mathbb{E}[Y] &= \theta\alpha_0 + (1 + \theta\alpha_1)\mathbb{E}[U_1] + (1 + \theta\alpha_2)\mathbb{E}[U_2], \\
\mathbb{E}[Z] &= \alpha_0 + \alpha_1\mathbb{E}[U_1] + \alpha_2\mathbb{E}[U_2].
\end{aligned}$$

### *S9.3.3 Partial identification in the presence of classical measurement error in the outcome*

**Theorem.** *Suppose the following conditions hold:*

- Rank preservation: $Y(A) = Y(0) + \theta A$, $\theta$ *constant.*
- Exposure-NCO independence given counterfactual outcome: $Z \perp\!\!\!\perp A | Y(0)$.
- NCO model: $Z = \beta_0^* + \beta_1^* Y(0) + \varepsilon$, $\varepsilon \perp\!\!\!\perp (A, Y(0)), \mathbb{E}[\varepsilon] = 0$.
- Classical measurement error: $Y = Y(A) + U$, $U \perp\!\!\!\perp (A, Y(0), Z)$, $\mathbb{E}[U] = 0$.

*Then,*

$$\theta \in \left[\hat{\theta}, \hat{\theta}\left(1 - R^2 \frac{1}{1 - \mathrm{Cor}^2(A,Y)}\right) + R^2 \frac{\mathrm{Var}(Y)}{\mathrm{Cov}(A,Y)}\left(1 - \frac{1}{1 - \mathrm{Cor}^2(A,Y)}\right)\right],$$

*where* $R^2 = 1 - \mathbb{E}[\mathrm{Var}(Y|A)]/\mathrm{Var}(Y)$ *is the proportion of variance of* $Y$ *explained by* $A$*, and* $\hat{\theta} = -\hat{\beta}_1/\hat{\beta}_2$ *and* $\hat{\beta}_1$ *and* $\hat{\beta}_1$ *are the ordinary least squares coefficients for* $A$ *and* $Y$ *in a linear regression of* $Z$ *on* $A$ *and* $Y$*.*

*Proof.* We have that

$$\begin{aligned} Z &= \beta_0^* + \beta_1^* Y(0) + \varepsilon && \text{(by NCO model)} \\ &= \beta_0^* + \beta_1^*(Y(A) - \theta A) + \varepsilon && \text{(by rank preservation)} \\ &= \beta_0^* + \beta_1^*(Y - U - \theta A) + \varepsilon && \text{(under classical measurement error)} \\ &= \beta_0^* + \beta_1^* Y - \beta_1^* U - \beta_1^* \theta A + \varepsilon, \end{aligned}$$

where $\varepsilon \perp\!\!\!\perp (Y, A, U)$ (since $U \perp\!\!\!\perp \varepsilon | (A, Y(0))$ and $\varepsilon \perp\!\!\!\perp (A, Y(0))$, so that $\varepsilon \perp\!\!\!\perp (Y(0), A, U)$).

Now, let

$$\begin{aligned} \hat{\beta}_1 &= \frac{\mathrm{Cov}(A,Z)\mathrm{Var}(Y) - \mathrm{Cov}(Y,Z)\mathrm{Cov}(A,Y)}{\mathrm{Var}(A)\mathrm{Var}(Y) - \mathrm{Cov}^2(A,Y)}, \\ \hat{\beta}_2 &= \frac{\mathrm{Cov}(Y,Z)\mathrm{Var}(A) - \mathrm{Cov}(A,Z)\mathrm{Cov}(A,Y)}{\mathrm{Var}(A)\mathrm{Var}(Y) - \mathrm{Cov}^2(A,Y)}, \end{aligned}$$

the ordinary least squares coefficients in a linear regression of $Z$ on $A$ and $Y$. We have

$$\begin{aligned} \mathrm{Cov}(A,Z) &= \beta_1^*(\mathrm{Cov}(A,Y) - \theta \mathrm{Var}(A)), \\ \mathrm{Cov}(Y,Z) &= \beta_1^*(\mathrm{Var}(Y) - \mathrm{Var}(U) - \theta \mathrm{Cov}(A,Y)), \end{aligned}$$

so that

$$\begin{aligned}
\hat{\beta}_1 &= \beta_1^* \left( \frac{\mathrm{Cov}(A,Y)\mathrm{Var}(U)}{\mathrm{Var}(A)\mathrm{Var}(Y) - \mathrm{Cov}^2(A,Y)} - \theta \right), \\
\hat{\beta}_2 &= \beta_1^* \left( 1 - \frac{\mathrm{Var}(A)\mathrm{Var}(U)}{\mathrm{Var}(A)\mathrm{Var}(Y) - \mathrm{Cov}^2(A,Y)} \right)
\end{aligned}$$

and in turn

$$\begin{aligned}
\hat{\theta} &= -\frac{\hat{\beta}_1}{\hat{\beta}_2} \\
&= -\frac{\mathrm{Cov}(A,Y)\mathrm{Var}(U) - \theta(\mathrm{Var}(A)\mathrm{Var}(Y) - \mathrm{Cov}^2(A,Y))}{\mathrm{Var}(A)\mathrm{Var}(U) - (\mathrm{Var}(A)\mathrm{Var}(Y) - \mathrm{Cov}^2(A,Y))}, \\
\theta &= \hat{\theta}\left( 1 - \frac{\mathrm{Var}(A)\mathrm{Var}(U)}{\mathrm{Var}(A)\mathrm{Var}(Y) - \mathrm{Cov}^2(A,Y)} \right) + \frac{\mathrm{Cov}(A,Y)\mathrm{Var}(U)}{\mathrm{Var}(A)\mathrm{Var}(Y) - \mathrm{Cov}^2(A,Y)} \\
&= \hat{\theta}\left( 1 - \frac{\mathrm{Var}(U)}{\mathrm{Var}(Y)} \frac{1}{1 - \mathrm{Cor}^2(A,Y)} \right) \\
&\qquad + \frac{\mathrm{Var}(U)}{\mathrm{Var}(Y)} \frac{\mathrm{Var}(Y)}{\mathrm{Cov}(A,Y)} \left( 1 - \frac{1}{1 - \mathrm{Cor}^2(A,Y)} \right).
\end{aligned}$$

By the law of total (conditional) variance,

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathbb{E}[\mathrm{Var}(Y|A)] + \mathrm{Var}(\mathbb{E}[Y|A]) \\
&= \mathbb{E}[\mathrm{Var}(Y|A, Y(0))|A] + \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y|A, Y(0)])|A] + \mathrm{Var}(\mathbb{E}[Y|A]) \\
&= \mathrm{Var}(U) + \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y|A, Y(0)])|A] + \mathrm{Var}(\mathbb{E}[Y|A]).
\end{aligned}$$

Now, define $R^2 = (\mathrm{Var}(Y) - \mathbb{E}[\mathrm{Var}(Y|A)])/\mathrm{Var}(Y)$, the proportion of variance of $Y$ explained by $A$ and observe that

$$R^2 \geq \frac{\mathrm{Var}(U)}{\mathrm{Var}(Y)} \geq 0.$$

Next, define

$$\tilde{\theta}(\lambda) = \hat{\theta}\left( 1 - \lambda \frac{1}{1 - \mathrm{Cor}^2(A,Y)} \right) + \lambda \frac{\mathrm{Var}(Y)}{\mathrm{Cov}(A,Y)} \left( 1 - \frac{1}{1 - \mathrm{Cor}^2(A,Y)} \right)$$

and note that, because the first derivative of $\tilde{\theta}$ is invariant to changes in $\lambda$, $\tilde{\theta}$ is monotonic. Hence

$$\theta \in [\tilde{\theta}(0), \tilde{\theta}(R^2)].$$

□

## References

DeGroot, M. and M. Schervish (2012): *Probability and Statistics*, Boston: Pearson, 4th edition.

# 10

# Identification of causal effects in case-control studies

Bas B. L. Penning de Vries
Rolf H. H. Groenwold

*BMC Medical Research Methodology* [Accepted for publication]

Identification of causal effects in case-control studies

## Abstract

Case-control designs are an important tool in contrasting the effects of well-defined treatments. In this paper, we reconsider classical concepts, assumptions and principles and explore when the results of case-control studies can be endowed a causal interpretation. Our focus is on identification of target causal quantities, or estimands. We cover various estimands relating to intention-to-treat or per-protocol effects for popular sampling schemes (case-base, survivor, and risk-set sampling), each with and without matching. Our approach may inform future research on different estimands, other variations of the case-control design or settings with additional complexities.

## 10.1 Introduction

In causal inference, it is important that the causal question of interest is unambiguously articulated (Hernán and Robins, 2020). The causal question should dictate, and therefore be at the start of, investigation. When the target causal quantity, the estimand, is made explicit, one can start to question how it relates to the available data distribution and, as such, form a basis for estimation with finite samples from this distribution.

The counterfactual framework offers a language rich enough to articulate a wide variety of causal claims that can be expressed as what-if statements (Hernán and Robins, 2020). Another, albeit closely related, approach to causal inference is target trial emulation, an explicit effort to mitigate departures from a study (the 'target trial') that, if carried out, would enable one to readily answer the causal what-if question of interest (Hernán and Robins, 2016). While it may be too impractical or unethical to implement, making explicit what a target trial looks like has particular value in communicating the inferential goal and offers a reference against which to compare studies that have been or are to be conducted.

The counterfactual framework and emulation approach have become increasingly popular in observational cohort studies. Case-control studies, however, have not yet enjoyed this trend. A notable exception is given by Dickerman et al. (2020), who recently outlined an application of trial emulation with case-control designs to statin use and colorectal cancer.

In this paper, we give an overview of how observational data obtained with case-control designs can be used to identify a number of causal estimands and, in doing so, recast historical case-control concepts, assumptions and principles in a modern and formal framework.

## 10.2 Preliminaries

### *10.2.1 Identification versus estimation*

An estimand is said to be identifiable if the distribution of the available data is compatible with exactly one value of the estimand, or therefore, if the estimand can be expressed as a function of the available data distribution. Identification forms a basis for estimation with finite samples from this distribution (Petersen and Van der Laan, 2014). Once the estimand has been made explicit and an identifiability expression established, estimation is a purely statistical task. While the expression will often naturally translate into a plug-in estimator, there is, however, generally more than one way to translate an identifiability result into an estimator and different estimators may have important differences in their statistical properties. Here, our focus is on identification, so that the purely statistical issues of the next step in causal inference, estimation, can be momentarily put aside.

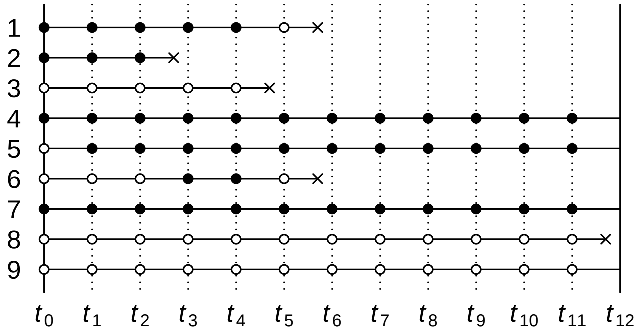
### *10.2.2 Case-control study nested in cohort study*

To facilitate understanding, it is useful to consider every case-control study as being "nested" within a cohort study. A case-control study is effectively a cohort study with missingness governed by the control sampling scheme. Therefore, when the observed data distribution of a case-control study is compatible with exactly one value of a given estimand, then so is the available or observed data distribution of the underlying cohort study. In other words, identifiability of an estimand with a case-control study implies identifiability of the estimand with the cohort study within which it is nested. The converse is not evident and in fact may not be true. In this paper, the focus is on sets of conditions or assumptions that are sufficient for identifiability in case-control studies.

### *10.2.3 Set-up of underlying cohort study*

Consider a time-varying exposure $A_k$ that can take one of two levels, 0 or 1, at $K$ successive time points $t_k$ ($k = 0, 1, ..., K - 1$), where $t_0$ denotes baseline (cohort

**Figure 10.1:** Illustration of possible courses of follow-up of an individual for a study with baseline $t_0$ and administrative study end $t_{12}$.



Solid bullets indicate 'exposed'; empty bullets indicate 'not exposed'. The incident event of interest is represented by a cross.

entry or time zero). Study participants are followed over time until they sustain the event of interest or the administrative study end $t_K$, whichever comes first. We denote by $T$ the time elapsed from baseline until the event of interest and let $Y_k = I(T < t_k)$ indicate whether the event has occurred by $t_k$. The lengths between the time points are typically fixed at a constant (e.g., of one day, week, or month). Figure 10.1 depicts twelve equally spaced time points over, say, twelve months with several possible courses of follow-up of an individual. As the figure illustrates, individuals can switch between exposure levels during follow-up, as in any truly observational study. Apart from exposure and outcome data, we also consider a (vector of) covariate(s) $L_k$, which describes time-fixed individual characteristics or time-varying characteristics typically relating to a time window just before exposure or non-exposure at $t_k$, $k = 0, 1, ..., K-1$.

### *10.2.4 Causal contrasts*

Although there are many possible contrasts, particularly with time-varying exposures, for simplicity we consider only two pairs of mutually exclusive interventions: (1) setting baseline exposure $A_0$ to 1 versus 0; and (2) setting all of $A_0, A_1, ..., A_{K-1}$ to 1 ('always exposed') versus all to 0 ('never exposed'). For $a = 0, 1$, we let counterfactual outcome $Y_k(a)$ indicate whether the event has occurred by $t_k$ under the baseline-only intervention that sets $A_0$ to $a$. By convention, we write $\overline{1} = (1, 1, ..., 1)$ and $\overline{0} = (0, 0, ..., 0)$, and let $Y_k(\overline{1})$ and $Y_k(\overline{0})$ indicate whether the event has occurred by $t_k$ under the intervention that sets $(A_0, A_1, ..., A_{K-1})$ all to 1 and all to 0, respectively. Further details about the notation and set-up are given in Supplementary Appendix S10.1.

### *10.2.5 Case-control sampling*

The fact that each time-specific exposure variable can take only one value per time point means that at most one counterfactual outcome can be observed per individual. This type of missingness is common to all studies. Relative to the cohort studies within which they are nested, case-control studies have additional missingness, which is governed by the control sampling scheme. In this paper, we focus on three well-known sampling schemes: case-base sampling, survivor sampling, and risk-set sampling. The next sections give an overview of conditions under which intention-to-treat and always-versus-never-exposed per-protocol effects can be identified with the data that are observed under these sampling schemes.

## 10.3 Case-control studies without matching

Table 10.1 summarises a number of identification results for case-control studies without matching. More formal statements and proofs are given in Supplementary Appendix S10.2. In all case-control studies that we consider in this section, cases are compared with controls with regard to their exposure status via an odds ratio, even when an effect measure other than the odds ratio is targeted. An individual qualifies as a case if and only if they sustain the event of interest by the administrative study end (i.e., $Y_K = 1$) and adhered to one of the protocols of interest until the time of the incident event. In Figure 10.1, the individual represented by row 1 is therefore regarded as a case (an exposed case in particular) in our investigation of intention-to-treat effects but not in that of per-protocol effects. Whether an individual (also) serves as a control depends on the control sampling scheme.

### *10.3.1 Case-base sampling*

The first result in Table 10.1 describes how to identify the intention-to-treat effect as quantified by the marginal risk ratio

$$\frac{\Pr(Y_K(1) = 1)}{\Pr(Y_K(0) = 1)}$$

under case-base sampling. (For identification of a conditional risk ratio, see Theorem 10.2 of Supplementary Appendix S10.2.) Case-base sampling, also known as case-cohort sampling, means that no individual who is at risk at baseline of sustaining the event of interest is precluded from selection as a control. Selection as a control, $S$, is further assumed independent of baseline covariate $L_0$ and exposure $A_0$. Selecting controls from survivors only (e.g., rows 4, 5, 7 and 9 in Figure 10.1) violates this assumption when survival depends on $L_0$ or $A_0$.

To account for baseline confounding, inverse probability weights could be derived from control data according to

$$W = \frac{A_0}{\Pr(A_0 = 1|L_0, S = 1)} + \frac{1 - A_0}{1 - \Pr(A_0 = 1|L_0, S = 1)}. \tag{10.1}$$

We then compute the odds of baseline exposure among cases and among controls in the pseudopopulation that is obtained by weighting everyone by subject-specific values of $W$. The ratio of these odds coincides with the target risk ratio under the three key identifiability conditions of consistency, baseline conditional exchangeability and positivity (Hernán and Robins, 2020).

The identification result for case-base sampling suggests a plug-in estimator: replace all functionals of the theoretical data distribution with sample analogues. For example, to obtain the weight for an individual with baseline covariate level $l_0$, replace the theoretical propensity score $\Pr(A_0 = 1|L_0 = l_0, S = 1)$ with an estimate $\widehat{\Pr}(A_0 = 1|L_0 = l_0, S = 1)$ derived from a fitted model (e.g., a logistic regression model) that imposes parametric constraints on the distribution of $A_0$ given $L_0$ among the controls.

### *10.3.2 Survivor sampling*

With survivor (cumulative incidence or exclusive) sampling, a subject is eligible for selection as a control only if they reach the administrative study end event-free. To identify the conditional odds ratio of baseline exposure versus baseline non-exposure given $L_0$,

$$\frac{\text{Odds}(Y_K(1) = 1|L_0)}{\text{Odds}(Y_K(0) = 1|L_0)},$$

selection as a control, $S$, is assumed independent of baseline exposure $A_0$ given $L_0$ and survival until the end of study (i.e., $Y_K = 0$).

The directed acyclic graph (DAG) of Figure 10.2 is compatible with both survivor sampling and case-base sampling. For those well versed in DAGs, it is tempting to conclude from it that restricting the analysis to those included in the study, i.e., conditioning on study inclusion, would result in bias (or departure from identification), by way of collider stratification. Although conditioning on study inclusion may indeed induce an association between baseline exposure and unmeasured cause $U$ of $Y_K$ (within levels of $L_0$), it is important to recognise it need not result in bias (Westreich, 2012; Hughes et al., 2019).

In fact, as is shown in Supplementary Appendix S10.2, Theorem 10.3, the above odds ratio is identified by the ratio of the baseline exposure odds given $L_0$ among the cases versus controls, provided the key identifiability conditions of consistency, baseline conditional exchangeability, and positivity are met.

All estimands in Table 10.1 describe a marginal effect, except for the odds ratio, which is conditional on baseline covariates $L_0$. The corresponding marginal odds ratio

$$\frac{\text{Odds}(Y_K(1) = 1)}{\text{Odds}(Y_K(0) = 1)}$$

is not identifiable from the available data distribution under the stated assumptions (see remark to Theorem 10.3, Supplementary Appendix S10.2).

**Table 10.1**: Overview of (non-parametric) identification results for case-control studies without matching.

| Sampling scheme | Estimand | Assumptions | Identification strategy |
|---|---|---|---|
| Case-base | Risk ratio for intention-to-treat effect<br>$\frac{\Pr(Y_K(1)=1)}{\Pr(Y_K(0)=1)}$ | • Control selection $S$ independent of baseline covariates $L_0$ and exposure $A_0$<br>• Consistency<br>• Baseline exchangeability given $L_0$<br>• Positivity<br>(Theorem 10.1) | 1. Derive time-fixed IP weights $W$ from control data<br>2. Compute the baseline exposure odds among cases, weighted by $W$<br>3. Compute the baseline exposure odds among controls, weighted by $W$ |
| Survivor | Odds ratio for intention-to-treat effect<br>$\frac{\text{Odds}(Y_K(1)=1\|L_0)}{\text{Odds}(Y_K(0)=1\|L_0)}$ | • Control selection $S$ independent of baseline exposure $A_0$ given baseline covariates $L_0$ and survival until $t_K$ ($Y_K=0$)<br>• Consistency<br>• Baseline exchangeability given $L_0$<br>• Positivity<br>(Theorem 10.3) | 1. Derive the conditional baseline exposure odds given $L_0$ among cases<br>2. Derive the conditional baseline exposure odds given $L_0$ among controls<br>3. Take the ratio of the results of steps 1 and 2 |

Table 10.1 continued.

| Sampling scheme | Estimand | Assumptions | Identification strategy |
|---|---|---|---|
| Risk-set | Hazard ratio for intention-to-treat effect<br>$\frac{\Pr(Y_{k+1}(1) = 1 \mid Y_k(1) = 0)}{\Pr(Y_{k+1}(0) = 1 \mid Y_k(0) = 0)}$ | • Control selection $S_k$ independent of baseline covariates $L_0$ and exposure $A_0$ given eligibility at $t_k$ ($Y_k = 0$) with constant sampling probability among those eligible$^\dagger$<br>• Consistency<br>• Baseline exchangeability given $L_0$<br>• Positivity<br>• Constant counterfactual hazards<br>(Theorem 10.4) | 1. Derive time-fixed IP weights $W$ from control data<br>2. Compute baseline exposure odds among cases, weighted by $W$<br>3. Compute baseline exposure odds among controls, weighted by $W$ times $\sum_{k=0}^{K-1} S_k$, the number of times selected as a control<br>4. Take the ratio of the results of steps 2 and 3 |

Table 10.1 continued.

| Sampling scheme | Estimand | Assumptions | Identification strategy |
|---|---|---|---|
| | Hazard ratio for per-protocol effect<br>$\frac{\Pr(Y_{k+1}(\overline{1}) = 1 \mid Y_k(\overline{1}) = 0)}{\Pr(Y_{k+1}(\overline{0}) = 1 \mid Y_k(\overline{0}) = 0)}$ | • Control selection $S_k$ independent of covariate and exposure history up to $t_k$ given eligibility at $t_k$ ($Y_k = 0$) with constant sampling probability among those eligible$^{\dagger}$<br>• Consistency<br>• Sequential conditional exchangeability<br>• Positivity<br>• Constant counterfactual hazards<br>(Theorem 10.6) | 1. Derive time-varying IP weights $W_k$ from control data<br>2. Censor from time of protocol deviation<br>3. Compute (baseline) exposure odds among cases, weighted by those weights $W_k$ such that $Y_k = 0$ and $Y_{k+1} = 1$<br>4. Compute (baseline) exposure odds among all controls, weighted by $\sum_{k=0}^{K-1} W_k S_k$, the weighted number of times selected as a control<br>5. Take the ratio of the results of steps 3 and 4 |

However, approximate identifiability can be achieved by invoking the rare event assumption (or rare disease assumption), in which case the marginal odds ratio approximates the marginal risk ratio.

### *10.3.3 Risk-set sampling for intention-to-treat effect*

With risk-set (or incidence density) sampling, for all time windows $[t_k, t_{k+1})$, $k = 0, ..., K-1$, every subject who is event-free at $t_k$ is eligible for selection as a control for the period $[t_k, t_{k+1})$. This means that study participants may be selected as a control more than once.

Consider the intention-to-treat effect quantified by the marginal (discrete-time) hazard ratio (or rate ratio)

$$\frac{\Pr(Y_{k+1}(1) = 1 | Y_k(1) = 0)}{\Pr(Y_{k+1}(0) = 1 | Y_k(0) = 0)}.$$

(For identification of a conditional hazard ratio, see Theorem 10.5, Supplementary Appendix S10.2.) For identification of the above marginal hazard ratio under risk-set sampling, it is assumed that selection as a control between $t_k$ and $t_{k+1}$, $S_k$, is independent of the baseline covariates and exposure given eligibility at $t_k$ (i.e., $Y_k = 0$). It is also assumed that the sampling probability among those eligible, $\Pr(S_k = 1 | Y_k = 0)$, is constant across time windows $k = 0, ..., K-1$. To this end, it suffices that the marginal hazard $\Pr(Y_{k+1} = 1 | Y_k = 0)$ remains constant across time windows and that every $k$th sampling fraction $\Pr(S_k = 1)$ is equal, up to a proportionality constant, to the probability $\Pr(Y_{k+1} = 1, Y_k = 0)$ of an incident case in the $k$th window (see remark to Theorem 10.4, Supplementary Appendix S10.2). For practical purposes, this suggests sampling a fixed number of controls for every case from among the set of eligible individuals. To illustrate, consider Figure 10.1 and note first of all that the individual represented by row 1 trivially qualifies as a case, because the individual survived until the event occurred. Because the event was sustained between $t_5$ and $t_6$, the proposed sampling suggests selecting a fixed number of controls from among those who are eligible at $t_5$. Thus, rows (and only rows) 4 through 9 as well as row 1 itself in Figure 10.1 qualify for selection as a control for this case. Even though the individual of row 1 is a case, the individual may also be selected as a control when the individuals of row 2, 3 and 6 (but not 8) sustain the event.

Once cases and controls are selected, we can start to derive inverse probability weights $W$ according to equation (10.1). We then compute the odds of baseline exposure among cases in the pseudopopulation that is obtained by weighting everyone by $W$ and the odds of baseline exposure among controls weighted by $W$

multiplied by the number of times the individual was selected as a control. The ratio of these odds coincides with the target hazard ratio under the three key identifiability conditions of consistency, baseline conditional exchangeability and positivity together with the assumption that the hazards in the numerator and denominator of the causal hazard ratio are constant across the time windows.

### *10.3.4 Risk-set sampling for per-protocol effect*

For the per-protocol effect quantified by the (discrete-time) hazard ratio (or rate ratio)

$$\frac{\Pr(Y_{k+1}(\overline{1}) = 1 | Y_k(\overline{1}) = 0)}{\Pr(Y_{j+1}(\overline{0}) = 1 | Y_k(\overline{0}) = 0)},$$

eligibility again requires that the respective subject is event-free at $t_k$ (i.e., $Y_k = 0$). Selection as a control between $t_k$ and $t_{k+1}$, $S_k$, is further assumed independent of covariate and exposure history up to $t_k$ given eligibility at $t_k$ (but see Supplementary Appendix S10.2 for a slightly weaker assumption). As for the intention-to-treat effect, it is also assumed that the probability to be selected as a control $S_k$ given eligibility is constant across time windows. This assumption is guaranteed to hold if the marginal hazard $\Pr(Y_{k+1} = 1 | Y_k = 0)$ remains constant across time windows and that every $k$th sampling fraction $\Pr(S_k = 1)$ is equal, up to a proportionality constant, to the probability of an incident case in the $k$th window. Figure 10.1 shows five incident events yet only three qualify as a case (rows 2, 3 and 8) when it concerns per-protocol effects. When the first case emerges (row 2), all rows meet the eligibility criterion for selection as a control. When the second emerges, the individual of row 2, who fails to survive event-free until $t_4$, is precluded as a control. When the case of row 8 emerges, only the individuals of rows 4, 5, 7 and 9 are eligible as controls.

Once cases and controls are selected, we can start to derive time-varying inverse probability weights according to

$$W_k = \prod_{j=0}^{k} \left[ \frac{A_j}{\Pr(A_j = 1 | L_0, ..., L_j, A_0, ..., A_{j-1}, Y_j = 0, S_j = 1)} \right. \\ \left. + \frac{1 - A_j}{1 - \Pr(A_j = 1 | L_0, ..., L_j, A_0, ..., A_{j-1}, Y_j = 0, S_j = 1)} \right].$$

It is important to note that the weights are derived from control information but are nonetheless used to weight both cases and controls (Robins, 1999).

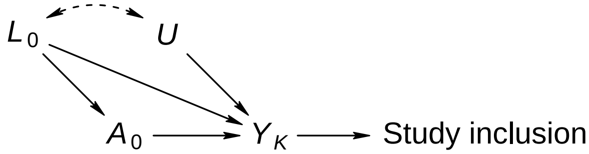
The denominators of the weights describe the propensity to switch exposure level. However, once the weights are derived, every subject is censored from the time that they fail to adhere to one of the protocols of interest for all downstream analysis. The uncensored exposure levels are therefore constant over time. We then compute the baseline exposure odds among cases, weighted by the weights $W_k$ corresponding to the interval $[t_k, t_{k+1})$ of the incident event (i.e., $Y_k = 0, Y_{k+1} = 1$), as well as the baseline exposure odds among controls, weighted by $\sum_{k=0}^{K-1} W_k S_k$, the weighted number of times selected as control. The ratio of these odds equals the target hazard ratio under the three key identifiability conditions of consistency, sequential conditional exchangeability, and positivity together with the assumption that hazards in the numerator and denominator of the causal hazard ratio for the per-protocol effect are constant across the time windows.

## 10.4 Case-control studies with matching

Table 10.2 gives an overview of identification results for case-control studies with exact pair matching. Formal statements and proofs are given in Supplementary Appendix S10.3, which also includes a generalisation of the results of Table 10.2 to exact 1-to-$M$ matching. While the focus in this section is on exact covariate matching, for partial matching we refer the reader to Supplementary Appendix S10.4, where we consider parametric identification by way of conditional logistic regression.

Pair matching involves assigning a single control exposure level, which we denote by $A'$, to every case. As for case-control studies without matching, in a case-control studies with matching an individual qualifies as a case if and only if they sustain the event of interest by the administrative study end (i.e., $Y_K = 1$)

**Figure 10.2:** Directed acyclic graph for a setting where inclusion (as case or control) into the case-control study with case-base or survivor sampling is determined by the outcome variable $Y_K$. $U$ represents an unknown or unobserved cause of $Y_K$. The dashed double-headed arrow represents an unmeasured or observed common cause.



and adhered to one of the protocols of interest until the time of the incident event. How a matched control exposure is assigned is encoded in the sampling scheme and the assumptions of Table 10.2. For example, for identification of the causal marginal risk ratio under case-base sampling, $A'$ is sampled from all study participants whose baseline covariate value matches that of the case, independently of the participants' baseline exposure value and whether they survive until the end of study. The matching is exact in the sense that the control exposure information is derived from an individual who has the same value for the baseline covariate as the case.

The identification strategy is the same for all results listed in Table 10.2. Only the case-control pairs $(A_0, A')$ with discordant exposure values (i.e., $(1, 0)$ or $(0, 1)$) are used. Under the stated sampling schemes and assumptions, the respective estimands are identified by the ratio of discordant pairs.

## 10.5 Discussion

This paper gives a formal account of how and when causal effects can be identified in case-control studies and, as such, underpins the case-control application of Dickerman et al. (2020). Like Dickerman et al., we believe that case-control studies should generally be regarded as being nested within cohort studies. This view emphasises that the threats to the validity of cohort studies should also be considered in case-control studies. For example, in case-control applications with risk-set sampling, researchers often consider the covariate and exposure status only at, or just before, the time of the event (for cases) or the time of sampling (for controls). However, where a cohort study would require information on baseline levels or the complete treatment and covariate history of participants, one should suspect that this holds for the nested case-control study too. To gain clarity, we encourage researchers to move away from using person-years, -weeks, or -days (rather than individuals) as the default units of inference (Hernán, 2015), and to realise that inadequately addressed deviations from a target trial may lead to bias (or departure from identifiability), regardless of whether the study that attempts to emulate it is a case-control or a cohort study (Dickerman et al., 2020).

What is meant by a cohort study differs between authors and contexts (Vandenbroucke and Pearce, 2012). The term 'cohort' may refer to either a 'dynamic population', or a 'fixed cohort', whose "membership is defined in a permanent fashion" and "determined by a single defining event and so becomes permanent" (Rothman et al., 2008). While it may sometimes be of interest to

**Table 10.2**: Overview of (non-parametric) identification results for case-control studies without matching.

| Sampling scheme | Estimand | Assumptions | Identification strategy |
|---|---|---|---|
| Case-base | Risk ratio for intention-to-treat effect<br>$$\frac{\Pr(Y_K(1)=1)}{\Pr(Y_K(0)=1)}$$ | • Matched control exposure $A'$ sampled from the baseline exposure levels of all subjects with same baseline covariate level $L_0$ as case, independently of the subjects' baseline exposure or survival status<br>• Consistency<br>• Baseline conditional exchangeability<br>• Positivity<br>• $\Pr(Y_K=1\|L_0=l,A_0=1)/\Pr(Y_K=1\|L_0=l,A_0=0)$ constant across levels $l$<br>(Theorem 10.7) | 1. Compute the frequency of discordant case-control pairs with $A_0=1$ and $A'=0$<br>2. Compute the frequency of discordant case-control pairs with $A_0=0$ and $A'=1$<br>3. Take the ratio of the results of steps 1 and 2 |

Table 10.2 continued.

| Sampling scheme | Estimand | Assumptions | Identification strategy |
|---|---|---|---|
| Survivor | Odds ratio for intention-to-treat effect<br>$\frac{\text{Odds}(Y_K(1) = 1 \mid L_0)}{\text{Odds}(Y_K(0) = 1 \mid L_0)}$ | • Matched control exposure $A'$ sampled from all the baseline exposure levels of all survivors ($Y_K = 0$) with same value for $L_0$ as case, independently of the subjects' baseline exposure<br>• Consistency<br>• Baseline conditional exchangeability<br>• Positivity<br>• $\text{Odds}(Y_K = 1 \mid L_0, A_0 = 1)/\text{Odds}(Y_K = 1 \mid L_0, A_0 = 0)$ constant across levels $l$ (Theorem 10.8) | (Same as identification strategy for case-base sampling) |

Table 10.2 continued.

| Sampling scheme | Estimand | Assumptions | Identification strategy |
|---|---|---|---|
| Risk-set | Hazard ratio for intention-to-treat effect<br>$\frac{\Pr(Y_{k+1}(1)=1\|L_0, Y_k(1)=0)}{\Pr(Y_{k+1}(0)=1\|L_0, Y_k(0)=0)}$ | • For a case with incident event in $[t_k, t_{k+1})$ (i.e., $Y_k=0, Y_{k+1}=1$), matched control exposure $A'$ sampled from the baseline exposure levels of all subjects that are event-free at $t_k$ ($Y_k=0$) and have the same value for $L_0$ as case. Sampling among these individuals is independent of baseline exposure or survival status<br>• Consistency<br>• Baseline conditional exchangeability<br>• Positivity<br>• $\Pr(Y_{k+1}=1\|L_0=l, A_0=1, Y_k=0)/\Pr(Y_{k+1}=1\|L_0=l, A_0=0, Y_k=0)$ constant across levels $k, l$<br>(Theorem 10.9) | (Same as identification strategy for case-base sampling) |

Table 10.2 continued.

| Sampling scheme | Estimand | Assumptions | Identification strategy |
|---|---|---|---|
| Risk-set | Hazard ratio for per-protocol effect[†] | • For a case with incident event in $[t+k, t_{k+1})$ (i.e., $Y_k = 0, Y_{k+1} = 1$), matched control exposure $A'$ sampled from the baseline exposure levels $A_0$ of all individuals who adhered to one of the protocols until $t_k$ (i.e., $A_0 = ... = A_k$) and have covariate history up to $t_k$. Sampling among these individuals is independent of baseline exposure or survival status<br>• Consistency<br>• Positivity<br>• $\Pr(Y_{k+1} = 1 \mid L_0, ..., L_k, A_0 = ... = A_k = 1, Y_k = 0)/$ $\Pr(Y_{k+1} = 1 \mid L_0, ..., L_k, A_0 = ... = A_k = 0, Y_k = 0)$ constant across levels $k$ and independent of $L_0, ..., L_k$ (Theorem 10.10) | (Same as identification strategy for case-base sampling) |

[†] $\dfrac{\Pr(Y_{k+1}(\overline{1}) = 1 \mid L_0, ..., L_k, A_0 = ... = A_k = 1, Y_k(\overline{1}) = 0)}{\Pr(Y_{k+1}(\overline{0}) = 1 \mid L_0, ..., L_k, A_0 = ... = A_k = 0, Y_k(\overline{0}) = 0)}$

ask what would have happened with a dynamic cohort (e.g., the residents of a country) had it been subjected to one treatment protocol versus another, the results in this paper relate to fixed cohorts.

Like the cohort studies within which they are (at least conceptually) nested, case-control studies require an explicit definition of time zero, the time at which a choice is to be made between treatment strategies or protocols of interest (Dickerman et al., 2020). Given a fixed cohort, time zero is generally determined by the defining event of the cohort (e.g., first diagnosis of a particular disease or having survived one year since diagnosis). This event may occur at different calendar times for different individuals. However, while a fixed cohort may be 'open' to new members relative to calendar time, it is always 'closed' along the time axis on which all subject-specific time zeros take a common point.

In this paper, time was regarded as discrete. Since we considered arbitrary intervals between time points and because, in real-world studies, time is never measured in a truly continuous fashion, this does not represent an important limitation for practical purposes. It is however important to note that the intervals between interventions and outcome assessments (in a target trial) are an intrinsic part of the estimand that lies at the start of investigation. Careful consideration of time intervals in the design of the conceptual target trial and of the actual cohort or case-control study is therefore warranted.

We emphasize that identification and estimation are distinct steps in causal inference. Although our focus was on the former, identifiability expressions often naturally translate into estimators. The task of finding the estimator with the most appealing statistical properties is not necessarily straightforward, however, and is beyond the scope of this paper.

We specifically studied two causal contrasts (i.e., pairs of interventions), one corresponding to intention-to-treat effects and the other to always-versus-never per-protocol effects of a time-varying exposure. There are of course many more causal contrasts, treatment regimes and estimands conceivable that could be of interest. We argue that also for these estimands, researchers should seek to establish identifiability before they select an estimator.

The conditions under which identifiability is to be sought for practical purposes may well include more constraints or obstacles to causal inference, such as additional missingness (e.g., outcome censoring) and measurement error, than we have considered here. While some of our results assume that hazards or hazard ratios remain constant over time, in many cases these are likely time-varying (Lefebvre et al., 2006; Guess, 2006). There are also more case-control designs (e.g., the case-crossover design) to consider. These additional complexities and designs are beyond the scope of this paper and represent an interesting direction for future research.

The case-control family of study designs is an important yet often misunderstood tool for identifying causal relations (Knol et al., 2008; Pearce, 2016; Mansournia et al., 2018; Labrecque et al., 2021). Although there is much to be learned, we believe that the modern arsenal for causal inference, which includes counterfactual thinking, is well-suited to make transparent for these classical epidemiological study designs what assumptions are sufficient or necessary to endow the study results with a causal interpretation and, in turn, help resolve or prevent misunderstanding.

## References

Dickerman, B. A., X. García-Albéniz, R. W. Logan, S. Denaxas, and M. A. Hernán (2020): "Emulating a target trial in case-control designs: an application to statins and colorectal cancer," *International Journal of Epidemiology*, 49, 1637–1646.

Guess, H. A. (2006): "Exposure-time-varying hazard function ratios in case-control studies of drug effects," *Pharmacoepidemiology and drug safety*, 15, 81–92.

Hernán, M. and J. Robins (2020): *Causal Inference: What If*, Boca Raton: Chapman & Hall/CRC.

Hernán, M. A. (2015): "Counterpoint: epidemiology to guide decision-making: moving away from practice-free research," *American journal of epidemiology*, 182, 834–839.

Hernán, M. A. and J. M. Robins (2016): "Using big data to emulate a target trial when a randomized trial is not available," *American journal of epidemiology*, 183, 758–764.

Hughes, R. A., J. Heron, J. A. Sterne, and K. Tilling (2019): "Accounting for missing data in statistical analyses: multiple imputation is not always the answer," *International journal of epidemiology*, 48, 1294–1304.

Knol, M. J., J. P. Vandenbroucke, P. Scott, and M. Egger (2008): "What do case-control studies estimate? survey of methods and assumptions in published case-control research," *American journal of epidemiology*, 168, 1073–1081.

Labrecque, J. A., M. M. Hunink, M. A. Ikram, and M. K. Ikram (2021): "Do case-control studies always estimate odds ratios?" *American journal of epidemiology*, 190, 318–321.

Lefebvre, G., J.-F. Angers, and L. Blais (2006): "Estimation of time-dependent rate ratios in case-control studies: comparison of two approaches for exposure assessment," *Pharmacoepidemiology and drug safety*, 15, 304–316.

Mansournia, M. A., N. P. Jewell, and S. Greenland (2018): "Case–control matching: effects, misconceptions, and recommendations," *European journal of epidemiology*, 33, 5–14.

Pearce, N. (2016): "Analysis of matched case-control studies," *BMJ*, 352.

Petersen, M. L. and M. J. van der Laan (2014): "Causal models and learning from data: integrating causal modeling and statistical estimation," *Epidemiology (Cambridge, Mass.)*, 25, 418–426.

Robins, J. M. (1999): "[choice as an alternative to control in observational studies]: comment," *Statistical Science*, 14, 281–293.

Rothman, K. J., S. Greenland, and T. L. Lash (2008): *Modern epidemiology*, Lippincott Williams & Wilkins, third edition edition.

Vandenbroucke, J. P. and N. Pearce (2012): "Incidence rates in dynamic populations," *International journal of epidemiology*, 41, 1472–1479.

Westreich, D. (2012): "Berkson's bias, selection bias, and missing data," *Epidemiology*, 23, 159–164.

## Supplementary Material

## S10.1 Notation and set-up

We will suppose that the interest lies with the effect of a time-varying exposure that can take one of two levels at any given time on a failure time outcome. In particular, we consider a strictly increasing sequence $(t_0, t_1, ..., t_K)$ of $K+1$ time points (with $t_{K+1} = -t_{-1} = +\infty$ for notational convenience). For $k = 0, 1, ..., K-1$, let $A_k$ denote the level of time-varying exposure of interest at $t_k$. We denote the history of any stochastic sequence $(X_0, X_1, ..., X_{K-1})$ up to and including $t_k$ by $\overline{X}_k = (X_0, X_1 ..., X_k)$ for $k = 0, 1, ..., K-1$ (and let $\overline{X} = \overline{X}_{K-1}$

and $\overline{X}_{-1} = 0$ for notational convenience). For example, $\overline{A} = (A_0, A_1, ..., A_{K-1})$. Denote by $T(\overline{a})$ the counterfactual time elapsed until the event of interest since $t_0$ that would have been realised had $\overline{A}$ been set to $\overline{a}$, and let $Y_k(\overline{a}) = I(T(\overline{a}) < t_k)$ for $k = 0, 1, ..., K$, where $I$ represents the indicator function. By convention, we stipulate that for all $k$, $Y_k(\overline{a})$ is invariant to the $k$th through $K-1$th elements of $\overline{a}$ (i.e., current survival status is not affected by future exposures). With slight abuse of notation, for $k = 0, 1..., K$, we let $Y_k(a_0)$ denote the outcome that would have been realised had (only) $A_0$ been set to $a_0$.

*Consistency*

For theorems about per-protocol effects, we assume consistency of the form: for $k = 1, ..., K$ and all $\overline{a}$, $Y_k(\overline{a}) = Y_k$ if $a_l = A_l$ for all $l = 0, ..., k-1$ such that $Y_l = 0$. For theorems about intention-to-treat effects, a weaker condition is sufficient and assumed: for $k = 1, ..., K$ and $a = 0, 1$, $Y_k(a) = Y_k$ if $a = A_0$. The assumption may be further relaxed for theorems in which the estimand does not involve $Y_k(a)$, $k < K$: for $a = 0, 1$, $Y_K(a) = Y_K$ if $a = A_0$.

*Conditional exchangeability*

We also consider a sequence of variables $\overline{L} = (L_0, L_1, ..., L_{K-1})$ that satisfies one of the following conditions:

$$\forall k, \forall \overline{a} : (Y_{k+1}(\overline{a}), ..., Y_K(\overline{a})) \perp\!\!\!\perp A_k | Y_k(\overline{a}) = 0, \overline{L}_k, \overline{A}_{k-1} = \overline{a}_{k-1},$$
(sequential conditional exchangeability, SCE)

where $\overline{a}_{k-1}$ is understood to represent the $(k-1)$th through $(K-1)$th elements of $\overline{a}$, or

$$\forall a_0 : (Y_1(a_0), ..., Y_K(a_0)) \perp\!\!\!\perp A_0 | L_0,$$
(baseline conditional exchangeability, BCE)

although sometimes a weaker form of BCE suffices: $\forall a_0 : Y_K(a_0) \perp\!\!\!\perp A_0 | L_0$.

*Positivity*

For the theorems that follow, we assume positivity to preclude division by zero and undefined conditional probabilities, so that the weights that we will encounter are finite and strictly greater than 1. The assumption can sometimes be relaxed if we are willing to interpolate or extrapolate under (parametric) modelling assumptions.

## S10.2 Identification results for non-matching strategies

*Intention-to-treat effect*

For simplicity, it is assumed below that the covariates are discrete. The results can however be extended to more general distributions.

**Theorem 10.1** (Case-base sampling for marginal intention-to-treat effect)**.** *Suppose BCE holds as well as*

$$\Pr(S=1|L_0,A_0)=\Pr(S=1)=\delta \tag{S1}$$

*for some* $\delta\in(0,1]$*. Then,*

$$\frac{\dfrac{\mathbb{E}\big[I(A_0=1)W|Y_K=1\big]}{\mathbb{E}\big[I(A_0=0)W|Y_K=1\big]}}{\dfrac{\mathbb{E}\big[I(A_0=1)W|S=1\big]}{\mathbb{E}\big[I(A_0=0)W|S=1\big]}}=\frac{\Pr(Y_K(1)=1)}{\Pr(Y_K(0)=1)},$$

*where*

$$W=\left.\frac{1}{\Pr(A_0=a|L_0,S=1)}\right|_{a=A_0},$$

*Proof.* First, observe that $\Pr(A_0=a|L_0,S=1)=\Pr(A_0=a|L_0)$ for $a=0,1$, because

$$\begin{aligned}\Pr(A_0=a|L_0,S=1)&=\frac{\Pr(S=1|L_0,A_0=a)\Pr(A_0=a|L_0)}{\Pr(S=1|L_0)}\\&=\frac{\delta}{\delta}\Pr(A_0=a|L_0) &&\text{(by S1)}\\&=\Pr(A_0=a|L_0)\end{aligned}$$

Hence,

$$W=\left.\frac{1}{\Pr(A_0=a|L_0)}\right|_{a=A_0}.$$

Now, consider the numerator of the left-hand side of the main equation in Theorem 10.1 and note that, because of the above, we have

$$\frac{\mathbb{E}\big[I(A_0=1)W|Y_K=1\big]}{\mathbb{E}\big[I(A_0=0)W|Y_K=1\big]}=\frac{\sum_{y=0}^{1}\mathbb{E}\big[I(A_0=1)WY_K|Y_K=y\big]\Pr(Y_K=y)}{\sum_{y=0}^{1}\mathbb{E}\big[I(A_0=0)WY_K|Y_K=y\big]\Pr(Y_K=y)}$$

$$= \frac{\mathbb{E}[I(A_0 = 1)WY_K]}{\mathbb{E}[I(A_0 = 0)WY_K]}$$

$$= \frac{\mathbb{E}[WY_K|A_0 = 1]\Pr(A_0 = 1)}{\mathbb{E}[WY_K|A_0 = 0]\Pr(A_0 = 0)},$$

where

$$\begin{aligned}
\mathbb{E}[WY_K|A_0 = a] &= \mathbb{E}\{\mathbb{E}[WY_K|L_0, A_0 = a]|A_0 = a\} \\
&= \sum_l \frac{\Pr(Y_K = 1|L_0 = l, A_0 = a)\Pr(L_0 = l|A_0 = a)}{\Pr(A_0 = a|L_0 = l)} \\
&= \sum_l \frac{\Pr(Y_K(a) = 1|L_0 = l, A_0 = a)\Pr(L_0 = l|A_0 = a)}{\Pr(A_0 = a|L_0 = l)} \\
&\hfill \text{(by consistency)} \\
&= \sum_l \frac{\Pr(Y_K(a) = 1|L_0 = l)\Pr(L_0 = l|A_0 = a)}{\Pr(A_0 = a|L_0 = l)} \\
&\hfill \text{(by baseline conditional exchangeability)} \\
&= \sum_l \frac{\Pr(Y_K(a) = 1|L_0 = l)\Pr(A_0 = a|L_0 = l)\Pr(L_0 = l)}{\Pr(A_0 = a|L_0 = l)\Pr(A_0 = a)} \\
&= \frac{1}{\Pr(A_0 = a)} \sum_l \Pr(Y_K(a) = 1, L_0 = l) \\
&= \frac{\Pr(Y_K(a) = 1)}{\Pr(A_0 = a)},
\end{aligned}$$

so that

$$\frac{\mathbb{E}[I(A_0 = 1)W|Y_K = 1]}{\mathbb{E}[I(A_0 = 0)W|Y_K = 1]} = \frac{\Pr(Y_K(1) = 1)}{\Pr(Y_K(0) = 1)}.$$

Next, consider the denominator of the left-hand side of the main equation in Theorem 10.1 and observe that

$$\frac{\mathbb{E}[I(A_0 = 1)W|S = 1]}{\mathbb{E}[I(A_0 = 0)W|S = 1]} = \frac{\mathbb{E}[I(A_0 = 1)WS]}{\mathbb{E}[I(A_0 = 0)WS]} = \frac{\mathbb{E}[WS|A_0 = 1]\Pr(A_0 = 1)}{\mathbb{E}[WS|A_0 = 0]\Pr(A_0 = 0)},$$

where

$$\begin{aligned}
\mathbb{E}[WS|A_0 = a] &= \mathbb{E}\{\mathbb{E}[WS|L_0, A_0 = a]|A_0 = a\} \\
&= \sum_l \frac{\Pr(S = 1|L_0, A_0 = a)\Pr(L_0 = l|A_0 = a)}{\Pr(A_0 = a|L_0 = l)}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_l \frac{\delta \Pr(L_0 = l | A_0 = a)}{\Pr(A_0 = a | L_0 = l)} && \text{(by S1)} \\
&= \frac{\delta}{\Pr(A_0 = a)} \sum_l \Pr(L_0 = l) \\
&= \frac{\delta}{\Pr(A_0 = a)},
\end{aligned}
$$

so that

$$\frac{\mathbb{E}\big[I(A_0 = 1)W | S = 1\big]}{\mathbb{E}\big[I(A_0 = 0)W | S = 1\big]} = 1.$$

It follows that

$$\frac{\dfrac{\mathbb{E}\big[I(A_0 = 1)W | Y_K = 1\big]}{\mathbb{E}\big[I(A_0 = 0)W | Y_K = 1\big]}}{\dfrac{\mathbb{E}\big[I(A_0 = 1)W | S = 1\big]}{\mathbb{E}\big[I(A_0 = 0)W | S = 1\big]}} = \frac{\Pr(Y_K(1) = 1)}{\Pr(Y_K(0) = 1)}.$$

□

**Theorem 10.2** (Case-base sampling for conditional intention-to-treat effect)**.** *Suppose BCE hold as well as S1, or the weaker version* $\Pr(S = 1 | L_0, A_0) = \Pr(S = 1 | L_0) = \delta_{L_0} \in (0, 1]$. *Then,*

$$\frac{\dfrac{\mathbb{E}\big[I(A_0 = 1) | L_0, Y_K = 1\big]}{\mathbb{E}\big[I(A_0 = 0) | L_0, Y_K = 1\big]}}{\dfrac{\mathbb{E}\big[I(A_0 = 1) | L_0, S = 1\big]}{\mathbb{E}\big[I(A_0 = 0) | L_0, S = 1\big]}} = \frac{\Pr(Y_K(1) = 1 | L_0)}{\Pr(Y_K(0) = 1 | L_0)}.$$

*Proof.* We have

$$
\begin{aligned}
&\frac{\mathbb{E}\big[I(A_0 = 1) | L_0, Y_K = 1\big]}{\mathbb{E}\big[I(A_0 = 0) | L_0, Y_K = 1\big]} \\
&\quad = \frac{\sum_{y=0}^{1} \mathbb{E}\big[I(A_0 = 1) Y_K | L_0, Y_K = y\big] \Pr(Y_K = y | L_0)}{\sum_{y=0}^{1} \mathbb{E}\big[I(A_0 = 0) Y_K | L_0, Y_K = y\big] \Pr(Y_K = y | L_0)} \\
&\quad = \frac{\mathbb{E}\big[I(A_0 = 1) Y_K | L_0\big]}{\mathbb{E}\big[I(A_0 = 0) Y_K | L_0\big]} \\
&\quad = \frac{\mathbb{E}\big[Y_K | L_0, A_0 = 1\big] \Pr(A_0 = 1 | L_0)}{\mathbb{E}\big[Y_K | L_0, A_0 = 0\big] \Pr(A_0 = 0 | L_0)}
\end{aligned}
$$

$$= \frac{\mathbb{E}[Y_K(1)|L_0, A_0 = 1] \Pr(A_0 = 1|L_0)}{\mathbb{E}[Y_K(0)|L_0, A_0 = 0] \Pr(A_0 = 0|L_0)} \qquad \text{(by consistency)}$$

$$= \frac{\mathbb{E}[Y_K(1)|L_0] \Pr(A_0 = 1|L_0)}{\mathbb{E}[Y_K(0)|L_0] \Pr(A_0 = 0|L_0)}. \quad \text{(by baseline conditional exchangeability)}$$

Also,

$$\begin{aligned}
\frac{\mathbb{E}[I(A_0 = 1)|L_0, S = 1]}{\mathbb{E}[I(A_0 = 0)|L_0, S = 1]} &= \frac{\mathbb{E}[I(A_0 = 1)S|L_0]}{\mathbb{E}[I(A_0 = 0)S|L_0]} \\
&= \frac{\mathbb{E}[S|L_0, A_0 = 1] \Pr(A_0 = 1|L_0)}{\mathbb{E}[S|L_0, A_0 = 0] \Pr(A_0 = 0|L_0)} \\
&= \frac{\delta_{L_0} \Pr(A_0 = 1|L_0)}{\delta_{L_0} \Pr(A_0 = 0|L_0)} \\
&\text{(under the assumption that } \Pr(S = 1|L_0, A_0) = \Pr(S = 1|L_0) = \delta_{L_0} \in (0, 1]) \\
&= \frac{\Pr(A_0 = 1|L_0)}{\Pr(A_0 = 0|L_0)}.
\end{aligned}$$

It immediately follows that

$$\frac{\dfrac{\mathbb{E}[I(A_0 = 1)|L_0, Y_K = 1]}{\mathbb{E}[I(A_0 = 0)|L_0, Y_K = 1]}}{\dfrac{\mathbb{E}[I(A_0 = 1)|L_0, S = 1]}{\mathbb{E}[I(A_0 = 0)|L_0, S = 1]}} = \frac{\Pr(Y_K(1) = 1|L_0)}{\Pr(Y_K(0) = 1|L_0)}.$$

□

**Corollary 10.1.** *If in addition to the conditions of Theorem 10.2,*

$$\frac{\Pr(Y_K = 1|L_0 = l, A_0 = 1)}{\Pr(Y_K = 1|L_0 = l, A_0 = 0)} = \theta \qquad \text{(homogeneity condition H1)}$$

*for all $l$ and some constant $\theta$, then*

$$\frac{\dfrac{\mathbb{E}[I(A_0 = 1)|L_0, Y_K = 1]}{\mathbb{E}[I(A_0 = 0)|L_0, Y_K = 1]}}{\dfrac{\mathbb{E}[I(A_0 = 1)|L_0, S = 1]}{\mathbb{E}[I(A_0 = 0)|L_0, S = 1]}} = \frac{\Pr(Y_K(1) = 1)}{\Pr(Y_K(0) = 1)},$$

*because of the collapsibility of the risk ratio.*

**Theorem 10.3** (Survivor sampling for conditional intention-to-treat effect)**.** *Suppose BCE holds as well as*

$$\Pr(S=1|L_0,A_0,Y_K)=\Pr(S=1|L_0,Y_K)=\delta_{L_0}\times(1-Y_K) \tag{S2}$$

*for some* $\delta_{L_0}\in(0,1]$. *Then,*

$$\frac{\dfrac{\mathbb{E}[I(A_0=1)|L_0,Y_K=1]}{\mathbb{E}[I(A_0=0)|L_0,Y_K=1]}}{\dfrac{\mathbb{E}[I(A_0=1)|L_0,S=1]}{\mathbb{E}[I(A_0=0)|L_0,S=1]}}=\frac{\text{Odds}(Y_K(1)=1|L_0)}{\text{Odds}(Y_K(0)=1|L_0)}.$$

*Proof.* First, consider the numerator of the left-hand side of the equation in Theorem 10.3 and observe

$$\begin{aligned}
\frac{\mathbb{E}[I(A_0=1)|L_0,Y_K=1]}{\mathbb{E}[I(A_0=0)|L_0,Y_K=1]}&=\frac{\Pr(Y_K=1|L_0,A_0=1)}{\Pr(Y_K=1|L_0,A_0=0)}\text{Odds}(A_0=1|L_0)\\
&=\frac{\Pr(Y_K(1)=1|L_0,A_0=1)}{\Pr(Y_K(1)=1|L_0,A_0=0)}\text{Odds}(A_0=1|L_0)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\text{(by consistency)}\\
&=\frac{\Pr(Y_K(1)=1|L_0)}{\Pr(Y_K(1)=1|L_0)}\text{Odds}(A_0=1|L_0).\\
&\qquad\qquad\text{(by baseline conditional exchangeability)}
\end{aligned}$$

Next, consider the denominator and observe that

$$\begin{aligned}
\frac{\mathbb{E}[I(A_0=1)|L_0,S=1]}{\mathbb{E}[I(A_0=0)|L_0,S=1]}&=\frac{\mathbb{E}[I(A_0=1)S|L_0]}{\mathbb{E}[I(A_0=0)S|L_0]}\\
&=\frac{\mathbb{E}[S|L_0,A_0=1]}{\mathbb{E}[S|L_0,A_0=0]}\text{Odds}(A_0=1|L_0)\\
&=\frac{\delta_{L_0}\Pr(Y_K=0|L_0,A_0=1)}{\delta_{L_0}\Pr(Y_K=0|L_0,A_0=0)}\text{Odds}(A_0=1|L_0)\ \text{(by S2)}\\
&=\frac{\Pr(Y_K(1)=0|L_0,A_0=1)}{\Pr(Y_K(0)=0|L_0,A_0=0)}\text{Odds}(A_0=1|L_0)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\text{(by consistency)}\\
&=\frac{\Pr(Y_K(1)=0|L_0)}{\Pr(Y_K(0)=0|L_0)}\text{Odds}(A_0=1|L_0).\\
&\qquad\qquad\text{(by baseline conditional exchangeability)}
\end{aligned}$$

It follows that

$$\frac{\dfrac{\mathbb{E}[I(A_0=1)|L_0,Y_K=1]}{\mathbb{E}[I(A_0=0)|L_0,Y_K=1]}}{\dfrac{\mathbb{E}[I(A_0=1)|L_0,S=1]}{\mathbb{E}[I(A_0=0)|L_0,S=1]}}=\frac{\mathrm{Odds}(Y_K(1)=1|L_0)}{\mathrm{Odds}(Y_K(0)=1|L_0)}.$$

□

**Remark to Theorem 10.3.** *Under BCE, the stronger version of S2,*

$$\Pr(S=1|L_0,A_0,Y_K)=\Pr(S=1|Y_K)=\delta\times(1-Y_K) \qquad \text{(S2}^*\text{)}$$

*for some* $\delta\in(0,1]$ *and with*

$$W=\left.\frac{1}{\Pr(A_0=a|L_0)}\right|_{a=A_0},$$

*we have*

$$\frac{\dfrac{\mathbb{E}[I(A_0=1)W|Y_K=1]}{\mathbb{E}[I(A_0=0)W|Y_K=1]}}{\dfrac{\mathbb{E}[I(A_0=1)W|S=1]}{\mathbb{E}[I(A_0=0)W|S=1]}}=\frac{\mathrm{Odds}(Y_K(1)=1)}{\mathrm{Odds}(Y_K(0)=1)} \qquad (10.2)$$

*(see proof below). However, from*

$$\begin{aligned}\Pr(A_0=a|L_0,S=1)&=\frac{\Pr(S=1|L_0,A_0=a)\Pr(A_0=a|L_0)}{\Pr(S=1|L_0)}\\&=\frac{\delta\Pr(Y_K=0|L_0,A_0=a)\Pr(A_0=a|L_0)}{\delta\Pr(Y_K=0|L_0)} \qquad \text{(by S2}^*\text{)}\\&=\Pr(A_0=a|L_0,Y_K=0),\end{aligned}$$

*it follows that the weights W above are not identified by*

$$\left.\frac{1}{\Pr(A_0=a|L_0,S=1)}\right|_{a=A_0}$$

*when* $Y_K \not\perp\!\!\!\perp A_0|L_0$. *(However,* $\Pr(A_0=a|L_0,S=1)$ *approximates* $\Pr(A_0=a|L_0)$ *under a rare event assumption.) In fact, the target marginal odds ratio is not identifiable, under BCE and S2* with unknown* $\delta$*, from the available data distribution, which is formed by the distribution of* $(L_0,A_0,Y_K,S)|(Y_K=1\vee S=1)$*. A proof is given below.*

*Proof of* (10.2) *under stated conditions.* As shown in the proof to Theorem 10.1,

$$\frac{\mathbb{E}[I(A_0 = 1)W|Y_K = 1]}{\mathbb{E}[I(A_0 = 0)W|Y_K = 1]} = \frac{\Pr(Y_K(1) = 1)}{\Pr(Y_K(0) = 1)}.$$

Now,

$$\frac{\mathbb{E}[I(A_0 = 1)W|S = 1]}{\mathbb{E}[I(A_0 = 0)W|S = 1]} = \frac{\mathbb{E}[I(A_0 = 1)WS]}{\mathbb{E}[I(A_0 = 0)WS]} = \frac{\mathbb{E}[WS|A_0 = 1]\Pr(A_0 = 1)}{\mathbb{E}[WS|A_0 = 0]\Pr(A_0 = 0)},$$

where

$$\begin{aligned}
\mathbb{E}[WS|A_0 = a] &= \mathbb{E}\{\mathbb{E}[WS|L_0, A_0 = a]|A_0 = a\} \\
&= \sum_l \frac{\Pr(S = 1|L_0, A_0 = a)\Pr(L_0 = l|A_0 = a)}{\Pr(A_0 = a|L_0 = l)} \\
&= \sum_l \frac{\delta \Pr(Y_K = 0|L_0 = l, A_0 = a)\Pr(L_0 = l|A_0 = a)}{\Pr(A_0 = a|L_0 = l)} \quad \text{(by S2}^*\text{)} \\
&= \frac{\delta}{\Pr(A_0 = a)} \sum_l \Pr(Y_K = 0|L_0 = l, A_0 = a)\Pr(L_0 = l) \\
&= \frac{\delta}{\Pr(A_0 = a)} \sum_l \Pr(Y_K(a) = 0|L_0 = l, A_0 = a)\Pr(L_0 = l) \\
&\qquad \text{(by consistency)} \\
&= \frac{\delta}{\Pr(A_0 = a)} \sum_l \Pr(Y_K(a) = 0, L_0 = l) \\
&\qquad \text{(by baseline conditional exchangeability)} \\
&= \frac{\delta \Pr(Y_K(a) = 0)}{\Pr(A_0 = a)},
\end{aligned}$$

so that

$$\frac{\mathbb{E}[I(A_0 = 1)W|S = 1]}{\mathbb{E}[I(A_0 = 0)W|S = 1]} = \frac{\Pr(Y_K(1) = 0)}{\Pr(Y_K(0) = 0)}$$

and, in turn,

$$\frac{\dfrac{\mathbb{E}[I(A_0 = 1)W|Y_K = 1]}{\mathbb{E}[I(A_0 = 0)W|Y_K = 1]}}{\dfrac{\mathbb{E}[I(A_0 = 1)W|S = 1]}{\mathbb{E}[I(A_0 = 0)W|S = 1]}} = \frac{\text{Odds}(Y_K(1) = 1)}{\text{Odds}(Y_K(0) = 1)}.$$

□

*Proof of nonidentifiability of target marginal odds ratio under stated conditions.* Consider two distributions of $(L_0, A_0, Y_K, S)$ satisfying S2$^*$, each characterised by the following conditionals:

$$\begin{aligned} Y_K &\sim \text{Bernoulli}(\alpha), \\ S|Y_K &\sim \text{Bernoulli}(\delta \times (1 - Y_K)), \\ L_0|Y_K, S \sim L_0|Y_K &\sim \text{Bernoulli}(5/10 - 2/10 \times Y_K), \\ A_0|L_0, Y_K, S \sim A_0|L_0, Y_K &\sim \text{Bernoulli}(3/10 + 2/10 \times L_0 + 3/10 \times Y_K). \end{aligned}$$

The parameter values of the distributions are given in the table below.

| Parameter | Distribution 1 | Distribution 2 |
|---|---|---|
| $\alpha$ | 1/10 | 2/10 |
| $\delta$ | 1/10 | 9/40 |

Now, for all $l, a, y, s \in \{0, 1\}$,

$$\begin{aligned} &\Pr(L_0 = l, A_0 = a, Y_K = y, S = s | Y_K = 1 \vee S = 1) \\ &\quad= \frac{\Pr(L_0 = l, A_0 = a, Y_K = y, S = s, Y_K = 1 \vee S = 1)}{\Pr(Y_K = 1 \wedge S = 0) + \Pr(Y_K = 0 \wedge S = 1) + \Pr(Y_K = 1 \wedge S = 1)} \\ &\quad= \frac{I(y = 1 \vee s = 1) \Pr(L_0 = l, A_0 = a, Y_K = y, S = s)}{\Pr(Y_K = 1) + \delta \Pr(Y_K = 0)} \\ &\quad= I(y = 1 \vee s = 1) \\ &\qquad\times \frac{\Pr(L_0 = l, A_0 = a | Y_K = y) \Pr(S = s | Y_K = y) \Pr(Y_K = y)}{\alpha + \delta(1 - \alpha)} \\ &\quad= \begin{cases} \Pr(L_0 = l, A_0 = a | Y_K = 0) \left(1 - \dfrac{\alpha}{\alpha + \delta(1 - \alpha)}\right) & \text{if } y = 0 \wedge s = 1, \\ \Pr(L_0 = l, A_0 = a | Y_K = 1) \dfrac{\alpha}{\alpha + \delta(1 - \alpha)} & \text{if } y = 1 \wedge s = 0, \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

where

$$\frac{\alpha}{\alpha + \delta(1 - \alpha)} = 10/19$$

under Distribution 1 and under Distribution 2. Hence, Distribution 1 and 2 imply the same available data distribution.

However, as we now show, the distributions imply different target marginal odds ratios. Since

$$\Pr(Y_K(a) = 1) = \sum_{l=0}^{1} \Pr(Y_K(a) = 1 | L_0 = l) \Pr(L_0 = l)$$

$$
\begin{aligned}
&= \sum_{l=0}^{1} \Pr(Y_K(a) = 1 | L_0 = l, A_0 = a) \Pr(L_0 = l) && \text{(by BCE)}\\
&= \sum_{l=0}^{1} \Pr(Y_K = 1 | L_0 = l, A_0 = a) \Pr(L_0 = l) && \text{(by consistency)}\\
&= \sum_{l=0}^{1} \frac{\Pr(L_0 = l, A_0 = a | Y_K = 1) \Pr(Y_K = 1)}{\Pr(L_0 = l, A_0 = a)}\\
&\qquad \times \sum_{y=0}^{1} \Pr(L_0 = l | Y_K = y) \Pr(Y_K = y)\\
&= \sum_{l=0}^{1} \left(1 + \frac{\Pr(L_0 = l, A_0 = a | Y_K = 0) \Pr(Y_K = 0)}{\Pr(L_0 = l, A_0 = a | Y_K = 1) \Pr(Y_K = 1)}\right)^{-1}\\
&\qquad \times \sum_{y=0}^{1} \Pr(L_0 = l | Y_K = y) \Pr(Y_K = y)
\end{aligned}
$$

for $a = 0, 1$, we have

$$
\begin{aligned}
\Pr(Y_K(1) = 1) &= \frac{5 + 2\alpha}{10 + (25/7)/\text{odds}(\alpha)} + \frac{5 - 2\alpha}{10 + (125/12)/\text{odds}(\alpha)} \quad \text{and}\\
\Pr(Y_K(0) = 1) &= \frac{5 + 2\alpha}{10 + (25/2)/\text{odds}(\alpha)} + \frac{5 - 2\alpha}{10 + (125/3)/\text{odds}(\alpha)},
\end{aligned}
$$

so that

$$
\frac{\text{Odds}(Y_K(1) = 1)}{\text{Odds}(Y_K(0) = 1)} = \begin{cases} \dfrac{587,791}{167,166} \approx 3.5 & \text{under Distribution 1,}\\ \dfrac{512,539}{148,789} \approx 3.4 & \text{under Distribution 2.} \end{cases}
$$

Hence, we found an available data distribution that is compatible with more than one value of the target marginal odds ratio. This concludes the proof. □

**Theorem 10.4** (Risk-set sampling for marginal intention-to-treat effect)**.** *Suppose BCE holds as well as*

$$
\Pr(S_k = 1 | L_0, A_0, Y_k) = \Pr(S_k = 1 | Y_k) = \delta \times (1 - Y_k), \tag{S3}
$$

*for some $\delta \in (0, 1]$. If*

$$
\Pr(Y_{k+1}(a) = 1 | Y_k(a) = 0) = \theta_a \tag{H2}
$$

*for* $a = 0, 1$ *and some constants* $\theta_0, \theta_1$*, then*

$$\frac{\dfrac{\mathbb{E}[I(A_0=1)W|Y_K=1]}{\mathbb{E}[I(A_0=0)W|Y_K=1]}}{\dfrac{\mathbb{E}[I(A_0=1)W\sum_{k=0}^{K-1}S_k]}{\mathbb{E}[I(A_0=0)W\sum_{k=0}^{K-1}S_k]}} = \frac{\Pr(Y_{k+1}(1)=1|Y_{k+1}(1)=0)}{\Pr(Y_{k+1}(0)=1|Y_{k+1}(0)=0)},$$

*where*

$$W = \left.\frac{1}{\Pr(A_0=a|L_0,S_0=1)}\right|_{a=A_0},$$

*Proof.* First, observe that $\Pr(A_0 = a|L_0, S_0 = 1) = \Pr(A_0 = a|L_0)$ for $a = 0, 1$, because

$$\begin{aligned}\Pr(A_0=a|L_0,S_0=1) &= \frac{\Pr(S_0=1|L_0,A_0=a)\Pr(A_0=a|L_0)}{\Pr(S_0=1|L_0)}\\ &= \frac{\delta}{\delta}\Pr(A_0=a|L_0) \qquad \text{(by S3)}\\ &= \Pr(A_0=a|L_0)\end{aligned}$$

Hence,

$$W = \left.\frac{1}{\Pr(A_0=a|L_0)}\right|_{a=A_0}.$$

For the numerator of the main result of Theorem 10.4, we thus have

$$\begin{aligned}\frac{\mathbb{E}[I(A_0=1)W|Y_K=1]}{\mathbb{E}[I(A_0=0)W|Y_K=1]} &= \frac{\mathbb{E}[I(A_0=1)WY_K]}{\mathbb{E}[I(A_0=0)WY_K]}\\ &= \frac{\mathbb{E}[WY_K|A_0=1]\Pr(A_0=1)}{\mathbb{E}[WY_K|A_0=0]\Pr(A_0=0)},\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}[WY_K|A_0=a] &= \mathbb{E}\{\mathbb{E}[WY_K|L_0,A_0=a]|A_0=a\}\\ &= \sum_l \frac{\Pr(Y_K=1|L_0=l,A_0=a)\Pr(L_0=l|A_0=a)}{\Pr(A_0=a|L_0=l)}\\ &= \sum_l \frac{\Pr(Y_K(a)=1|L_0=l,A_0=a)\Pr(L_0=l|A_0=a)}{\Pr(A_0=a|L_0=l)}\\ &\qquad \text{(by consistency)}\end{aligned}$$

$$= \sum_l \frac{\Pr(Y_K(a) = 1|L_0 = l)\Pr(L_0 = l|A_0 = a)}{\Pr(A_0 = a|L_0 = l)}$$

(by baseline conditional exchangeability)

$$= \sum_l \frac{\Pr(Y_K(a) = 1|L_0 = l)\Pr(A_0 = a|L_0 = l)\Pr(L_0 = l)}{\Pr(A_0 = a|L_0 = l)\Pr(A_0 = a)}$$

$$= \frac{1}{\Pr(A_0 = a)}\sum_l \Pr(Y_K(a) = 1, L_0 = l)$$

$$= \frac{\Pr(Y_K(a) = 1)}{\Pr(A_0 = a)},$$

so that

$$\begin{aligned}
\frac{\mathbb{E}[I(A_0 = 1)W|Y_K = 1]}{\mathbb{E}[I(A_0 = 0)W|Y_K = 1]} &= \frac{\Pr(Y_K(1) = 1)}{\Pr(Y_K(0) = 1)} \\
&= \frac{\sum_{k=0}^{K-1}\Pr(Y_{k+1}(1) = 1, Y_k(1) = 0)}{\sum_{k=0}^{K-1}\Pr(Y_{k+1}(0) = 1, Y_k(0) = 0)} \\
&= \frac{\sum_{k=0}^{K-1}\Pr(Y_{k+1}(1) = 1|Y_k(1) = 0)\Pr(Y_k(1) = 0)}{\sum_{k=0}^{K-1}\Pr(Y_{k+1}(0) = 1|Y_k(0) = 0)\Pr(Y_k(0) = 0)} \\
&= \frac{\sum_{k=0}^{K-1}\theta_1\Pr(Y_k(1) = 0)}{\sum_{k=0}^{K-1}\theta_0\Pr(Y_k(0) = 0)} && \text{(by H2)} \\
&= \frac{\theta_1}{\theta_0}\frac{\sum_{k=0}^{K-1}\Pr(Y_k(1) = 0)}{\sum_{k=0}^{K-1}\Pr(Y_k(0) = 0)}
\end{aligned}$$

For the denominator, we have

$$\frac{\mathbb{E}[I(A_0 = 1)W\sum_{k=0}^{K-1}S_k]}{\mathbb{E}[I(A_0 = 0)W\sum_{k=0}^{K-1}S_k]} = \frac{\mathbb{E}[W\sum_{k=0}^{K-1}S_k|A_0 = 1]\Pr(A_0 = 1)}{\mathbb{E}[W\sum_{k=0}^{K-1}S_k|A_0 = 0]\Pr(A_0 = 0)},$$

where

$$\begin{aligned}
&\mathbb{E}[W\textstyle\sum_{k=0}^{K-1}S_k|A_0 = a] \\
&\quad = \textstyle\sum_{k=0}^{K-1}\mathbb{E}\{\mathbb{E}[WS_k|L_0, A_0 = a]|A_0 = a\} \\
&\quad = \sum_{k=0}^{K-1}\sum_l \frac{\Pr(S_k = 1|L_0, A_0 = a)\Pr(L_0 = l|A_0 = a)}{\Pr(A_0 = a|L_0 = l)} \\
&\quad = \sum_{k=0}^{K-1}\sum_l \frac{\delta\Pr(Y_k = 0|L_0 = l, A_0 = a)\Pr(L_0 = l|A_0 = a)}{\Pr(A_0 = a|L_0 = l)} && \text{(by S3)}
\end{aligned}$$

$$= \sum_{k=0}^{K-1} \sum_{l} \frac{\delta \Pr(Y_k = 0 | L_0 = l, A_0 = a) \Pr(L_0 = l)}{\Pr(A_0 = a)}$$

$$= \sum_{k=0}^{K-1} \sum_{l} \frac{\delta \Pr(Y_k(a) = 0 | L_0 = l, A_0 = a) \Pr(L_0 = l)}{\Pr(A_0 = a)} \qquad \text{(by consistency)}$$

$$= \sum_{k=0}^{K-1} \sum_{l} \frac{\delta \Pr(Y_k(a) = 0 | L_0 = l) \Pr(L_0 = l)}{\Pr(A_0 = a)}$$

(by baseline conditional exchangeability)

$$= \frac{1}{\Pr(A_0 = a)} \sum_{k=0}^{K-1} \sum_{l} \delta \Pr(Y_k(a) = 0, L_0 = l)$$

$$= \frac{1}{\Pr(A_0 = a)} \sum_{k=0}^{K-1} \delta \Pr(Y_k(a) = 0),$$

so that

$$\frac{\mathbb{E}[I(A_0 = 1) W \sum_{k=0}^{K-1} S_k]}{\mathbb{E}[I(A_0 = 0) W \sum_{k=0}^{K-1} S_k]} = \frac{\sum_{k=0}^{K-1} \delta \Pr(Y_k(1) = 0)}{\sum_{k=0}^{K-1} \delta \Pr(Y_k(0) = 0)} = \frac{\sum_{k=0}^{K-1} \Pr(Y_k(1) = 0)}{\sum_{k=0}^{K-1} \Pr(Y_k(0) = 0)}.$$

It follows that

$$\frac{\dfrac{\mathbb{E}[I(A_0 = 1) W | Y_K = 1]}{\mathbb{E}[I(A_0 = 0) W | Y_K = 1]}}{\dfrac{\mathbb{E}[I(A_0 = 1) W \sum_{k=0}^{K-1} S_k]}{\mathbb{E}[I(A_0 = 0) W \sum_{k=0}^{K-1} S_k]}} = \frac{\Pr(Y_{k+1}(1) = 1 | Y_k(1) = 0)}{\Pr(Y_{k+1}(0) = 1 | Y_k(0) = 0)}.$$

□

**Remark to Theorem 10.4.** *Condition S3 holds if, for some constant $\delta_k^*$,*

$$\left.\begin{array}{c} \Pr(S_k = 1) = \delta_k^* \Pr(Y_{k+1} = 1, Y_k = 0), \\ S_k \perp\!\!\!\perp (L_0, A_0, \overline{Y}_k) | Y_k = 0, \\ \Pr(S_k = 1 | Y_k = 1) = 0. \end{array}\right\} \qquad \text{(S3}^*\text{)}$$

*The first requirement of S3* essentially means that the frequency of incident cases in the kth window is proportional to the frequency of controls selected in this window. Under S3*, S3 is met with $\delta = \delta_k^* \Pr(Y_{k+1} = 1 | Y_k = 0)$, because*

$$\Pr(S_k = 1 | L_0, A_0, \overline{Y}_k) = \Pr(S_k = 1 | Y_k)$$

$$
\begin{aligned}
&= \Pr(S_k = 1|Y_k = 0) \times (1 - Y_k) \\
&= \Pr(S_k = 1|Y_k = 0) \times (1 - Y_k) \\
&= \frac{\Pr(S_k = 1)}{\Pr(Y_k = 0)} \times (1 - Y_k) \\
&= \frac{\delta_k^* \Pr(Y_{k+1} = 1, Y_k = 0)}{\Pr(Y_k = 0)} \times (1 - Y_k) \\
&= \delta_k^* \Pr(Y_{k+1} = 1|Y_k = 0) \times (1 - Y_k).
\end{aligned}
$$

*Therefore, stipulating that $\delta_k^*$ is k-invariant is to state that* $\Pr(Y_{k+1} = 1|Y_k = 0)$ *is constant for $k = 0, ..., K - 1$.*

**Theorem 10.5** (Risk-set sampling for conditional intention-to-treat effect)**.** *Suppose BCE holds as well as S3, or the weaker version* $\Pr(S_k = 1|L_0, A_0, Y_k) = \Pr(S_k = 1|L_0, Y_k) = \delta_{L_0} \times (1 - Y_k),\ \delta_{L_0} \in (0, 1]$. *If*

$$\Pr(Y_{k+1}(a) = 1|L_0 = l, Y_k(a) = 0) = \theta_a \tag{H3}$$

*for $a = 0, 1$, all $l$ and some constants $\theta_0, \theta_1$, then*

$$\frac{\dfrac{\mathbb{E}[I(A_0 = 1)|L_0, Y_K = 1]}{\mathbb{E}[I(A_0 = 0)|L_0, Y_K = 1]}}{\dfrac{\mathbb{E}[I(A_0 = 1)\sum_{k=0}^{K-1} S_k|L_0]}{\mathbb{E}[I(A_0 = 0)\sum_{k=0}^{K-1} S_k|L_0]}} = \frac{\Pr(Y_{k+1}(1) = 1|L_0, Y_k(1) = 0)}{\Pr(Y_{k+1}(0) = 1|L_0, Y_k(0) = 0)}.$$

The proof to Theorem 10.5 is similar to that of Theorem 10.4 and therefore omitted.

### *Per-protocol effect*

In this subsection, an individual qualifies as a case if and only if $Y_K = 1$ and the subject adheres to the protocol that was assigned at baseline. For any study participant, let $S_k$ denote selection as a control for the period $[t_k, t_{k+1})$ and suppose $S_k$ satisfies

$$\left.\begin{array}{c}
S_k = 1 \Rightarrow Y_k = 0 \ \text{ with probability 1, and} \\
\Pr(S_k = 1|\overline{L}_k, \overline{A}_k, Y_k = 0) = \Pr(S_k = 1|\overline{A}_{k-1}, Y_k = 0) \ \text{ and} \\
\Pr(S_k = 1|\overline{A}_{k-1}, A_0 = ... = A_{k-1}, Y_k = 0) = \delta,
\end{array}\right\} \tag{S4}$$

for some $\delta \in (0, 1]$.

**Remark to Theorem 10.6.** *Condition S4 holds if, for some constant* $\delta_k^*$,

$$\left.\begin{array}{c}\Pr(S_k=1)=\delta_k^*\Pr(Y_{k+1}=1,Y_k=0,\forall j<k:A_j=A_0)\ \ \textit{and}\\ S_k \perp\!\!\!\perp (\overline{L}_k,\overline{A}_k,\overline{Y}_k)|(Y_k=0,\forall j<k:A_j=A_0)\ \ \textit{and}\\ S_k=1\Rightarrow(Y_k=0,\forall j<k:A_j=A_0)\ \ \textit{with probability 1.}\end{array}\right\}\qquad(\mathrm{S4}^*)$$

*The first requirement of S4*$^*$ *essentially means that the frequency of protocol-adherent incident cases in the kth window is proportional to the frequency of controls selected in this window. Under S4*$^*$, *S4 is met with* $\delta=\delta_k^*\Pr(Y_{k+1}=1|Y_k=0,\forall j<k:A_j=A_0)$, *because*

$$\begin{aligned}&\Pr(S_k=1|\overline{L}_k,\overline{A}_k,\overline{Y}_k)\\ &\quad=\Pr(S_k=1|Y_k=0,\forall j<k:A_j=A_0)\\ &\qquad\times(1-Y_k)\times I(\forall j<k:A_j=A_0)\\ &\quad=\frac{\Pr(S_k=1)}{\Pr(Y_k=0,\forall j<k:A_j=A_0)}\times(1-Y_k)\times I(\forall j<k:A_j=A_0)\\ &\quad=\frac{\delta_k^*\Pr(Y_{k+1}=1,Y_k=0,\forall j<k:A_j=A_0)}{\Pr(Y_k=0,\forall j<k:A_j=A_0)})\\ &\qquad\times(1-Y_k)\times I(\forall j<k:A_j=A_0)\\ &\quad=\delta_k^*\Pr(Y_{k+1}=1|Y_k=0,\forall j<k:A_j=A_0))\\ &\qquad\times(1-Y_k)\times I(\forall j<k:A_j=A_0).\end{aligned}$$

*Similarly, condition S4 holds if, for some constant* $\delta_k^{**}$,

$$\left.\begin{array}{c}\Pr(S_k=1)=\delta_k^{**}\Pr(Y_{k+1}=1,Y_k=0)\ \ \textit{and}\\ S_k \perp\!\!\!\perp (\overline{L}_k,\overline{A}_k,\overline{Y}_k)|(Y_k=0)\ \ \textit{and}\\ S_k=1\Rightarrow Y_k=0\ \ \textit{with probability 1,}\end{array}\right\}\qquad(\mathrm{S4}^{**})$$

*in which case,* $\delta=\delta_k^{**}\Pr(Y_{k+1}=1|Y_k=0)$, *because*

$$\begin{aligned}&\Pr(S_k=1|\overline{L}_k,\overline{A}_k,\overline{Y}_k)\\ &\quad=\Pr(S_k=1|Y_k=0)\times(1-Y_k)\\ &\quad=\frac{\Pr(S_k=1)}{\Pr(Y_k=0)}\times(1-Y_k)\\ &\quad=\frac{\delta_k^{**}\Pr(Y_{k+1}=1,Y_k=0)}{\Pr(Y_k=0)}\times(1-Y_k)\\ &\quad=\delta_k^{**}\Pr(Y_{k+1}=1|Y_k=0)\times(1-Y_k).\end{aligned}$$

**Theorem 10.6** (Risk-set sampling for marginal per-protocol effect)**.** *Suppose SCE and S4 hold. If*

$$\Pr(Y_{k+1}(\overline{a}) = 1 | Y_k(\overline{a}) = 0) = \theta_a \tag{H4}$$

*for* $a = 0, 1$ *and some constants* $\theta_0, \theta_1$, *then*

$$\frac{\dfrac{\mathbb{E}\left[\sum_{k=0}^{K-1} I(A_k = 1) W_k I(Y_{k+1} = 1, Y_k = 0) | Y_K = 1, (\forall j : Y_j = 0 \Rightarrow A_j = A_0)\right]}{\mathbb{E}\left[\sum_{k=0}^{K-1} I(A_k = 0) W_k I(Y_{k+1} = 1, Y_k = 0) | Y_K = 1, (\forall j : Y_j = 0 \Rightarrow A_j = A_0)\right]}}{\dfrac{\mathbb{E}\left[I(A_0 = 1) \sum_{k=0}^{K-1} W_k S_k | \forall j : Y_j = 0 \Rightarrow A_j = A_0\right]}{\mathbb{E}\left[I(A_0 = 0) \sum_{k=0}^{K-1} W_k S_k | \forall j : Y_j = 0 \Rightarrow A_j = A_0\right]}}$$
$$= \frac{\Pr(Y_{k+1}(\overline{1}) = 1 | Y_k(\overline{1}) = 0)}{\Pr(Y_{k+1}(\overline{0}) = 1 | Y_k(\overline{0}) = 0)},$$

*where*

$$W_k = \prod_{j=0}^{k} \frac{1}{\Pr(A_j = a_j | \overline{L}_j, \overline{A}_{j-1}, Y_j = 0, S_j = 1)} \bigg|_{a_j = A_j}.$$

*Proof.* First, observe that $\Pr(A_k = a' | \overline{L}_k, (\forall j < k : A_j = a), Y_k = 0, S_k = 1) = \Pr(A_k = a' | \overline{L}_k, (\forall j < k : A_j = a), Y_k = 0)$ for $a', a = 0, 1$, because

$$\begin{aligned}
&\Pr(A_k = a' | \overline{L}_k, (\forall j < k : A_j = a), Y_k = 0, S_k = 1) \\
&\quad = \frac{\Pr(S_k = 1 | \overline{L}_k, (\forall j < k : A_j = a), A_k = a', Y_k = 0) \times \Pr(A_k = a' | \overline{L}_k, (\forall j < k : A_j = a), Y_k = 0)}{\Pr(S_k = 1 | \overline{L}_k, (\forall j < k : A_j = a), Y_k = 0)} \\
&\quad = \frac{\delta}{\delta} \Pr(A_k = a' | \overline{L}_k, (\forall j < k : A_j = a), Y_k = 0). && \text{(by S4)}
\end{aligned}$$

Hence, if $\forall j < k : A_j = A_0$, then

$$W_k = \prod_{j=0}^{k} \frac{1}{\Pr(A_j = a_j | \overline{L}_j, \overline{A}_{j-1}, Y_j = 0)} \bigg|_{a_j = A_j}.$$

For the numerator of the main result of Theorem 10.6, we thus have

$$\frac{\mathbb{E}\left[\sum_{k=0}^{K-1} I(A_k = 1) W_k I(Y_{k+1} = 1, Y_k = 0) | Y_K = 1, (\forall j : Y_j = 0 \Rightarrow A_j = A_0)\right]}{\mathbb{E}\left[\sum_{k=0}^{K-1} I(A_k = 0) W_k I(Y_{k+1} = 1, Y_k = 0) | Y_K = 1, (\forall j : Y_j = 0 \Rightarrow A_j = A_0)\right]}$$

$$\frac{\mathbb{E}\left[\sum_{k=0}^{K-1} I(A_k = a) W_k I(Y_{k+1} = 1, Y_k = 0, \forall j \leq k : A_j = A_0)\right]}{\mathbb{E}\left[\sum_{k=0}^{K-1} I(A_k = a') W_k I(Y_{k+1} = 1, Y_k = 0, \forall j \leq k : A_j = A_0)\right]}$$

$$= \frac{\sum_{k=0}^{K-1} \mathbb{E}[W_k Y_{k+1}(1-Y_k) I(\forall j \le k : A_j = a)]}{\sum_{k=0}^{K-1} \mathbb{E}[W_k Y_{k+1}(1-Y_k) I(\forall j \le k : A_j = a')]}$$

$$= \frac{\displaystyle\sum_{k=0}^{K-1} \sum_{\bar{l}_k} \frac{\Pr(Y_{k+1} = 1, Y_k = 0, \forall j \le k : A_j = a, \overline{L}_k = \bar{l}_k)}{\prod_{j=0}^{k} \Pr(A_j = a | Y_j = 0, \overline{L}_k = \bar{l}_k, \forall i < j : A_i = a)}}{\displaystyle\sum_{k=0}^{K-1} \sum_{\bar{l}_k} \frac{\Pr(Y_{k+1} = 1, Y_k = 0, \forall j \le k : A_j = a', \overline{L}_k = \bar{l}_k)}{\prod_{j=0}^{k} \Pr(A_j = a' | Y_j = 0, \overline{L}_k = \bar{l}_k, \forall i < j : A_i = a')}},$$

where

$$\begin{aligned}
&\sum_{\bar{l}_k} \frac{\Pr(Y_{k+1} = 1, Y_k = 0, \forall j \le k : A_j = a, \overline{L}_k = \bar{l}_k)}{\prod_{j=0}^{k} \Pr(A_j = a | Y_j = 0, \overline{L}_k = \bar{l}_k, \forall i < j : A_i = a)} \\
&\quad = \sum_{\bar{l}_k} \Pr(Y_{k+1} = 1 | Y_k = 0, \overline{L}_k = \bar{l}_k, \forall j \le k : A_j = a) \\
&\qquad \times \Pr(L_k = l_k | Y_k = 0, \overline{L}_{k-1} = \bar{l}_{k-1}, \forall j < k : A_j = a) \\
&\qquad \times \prod_{j=0}^{k-1} \Pr(Y_{j+1} = 1 | Y_j = 0, \overline{L}_j = \bar{l}_j, \forall i \le j : A_i = a) \\
&\qquad \times \Pr(L_j = l_j | Y_j = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j : A_i = a) \\
&\quad = \sum_{\bar{l}_k} \Pr(Y_{k+1}(\overline{a}) = 1 | Y_k(\overline{a}) = 0, \overline{L}_k = \bar{l}_k, \forall j \le k : A_j = a) \\
&\qquad \times \Pr(L_k = l_k | Y_k(\overline{a}) = 0, \overline{L}_{k-1} = \bar{l}_{k-1}, \forall j < k : A_j = a) \\
&\qquad \times \prod_{j=0}^{k-1} \Pr(Y_{j+1}(\overline{a}) = 1 | Y_j(\overline{a}) = 0, \overline{L}_j = \bar{l}_j, \forall i \le j : A_i = a) \\
&\qquad \times \Pr(L_j = l_j | Y_j(\overline{a}) = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j : A_i = a) \\
&\hspace{20em} \text{(by consistency)} \\
&\quad = \sum_{\bar{l}_k} \Pr(Y_{k+1}(\overline{a}) = 1 | Y_k(\overline{a}) = 0, \overline{L}_k = \bar{l}_k, \forall j < k : A_j = a) \\
&\qquad \times \Pr(L_k = l_k | Y_k(\overline{a}) = 0, \overline{L}_{k-1} = \bar{l}_{k-1}, \forall j < k : A_j = a) \\
&\qquad \times \prod_{j=0}^{k-1} \Pr(Y_{j+1}(\overline{a}) = 1 | Y_j(\overline{a}) = 0, \overline{L}_j = \bar{l}_j, \forall i < j : A_i = a) \\
&\qquad \times \Pr(L_j = l_j | Y_j(\overline{a}) = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j : A_i = a) \\
&\hspace{12em} \text{(by sequential conditional exchangeability)}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\bar{l}_{k-1}} \Pr(Y_{k+1}(\overline{a}) = 1 | Y_k(\overline{a}) = 0, \overline{L}_{k-1} = \bar{l}_{k-1}, \forall j < k : A_j = a) \\
&\quad \times \prod_{j=0}^{k-1} \Pr(Y_{j+1}(\overline{a}) = 1 | Y_j(\overline{a}) = 0, \overline{L}_j = \bar{l}_j, \forall i < j : A_i = a) \\
&\quad \times \Pr(L_j = l_j | Y_j(\overline{a}) = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j : A_i = a) \\
&= \sum_{\bar{l}_{k-1}} \Pr(Y_{k+1}(\overline{a}) = 1, Y_k(\overline{a}) = 0 | Y_{k-1}(\overline{a}) = 0, \overline{L}_{k-1} = \bar{l}_{k-1}, \forall j < k : A_j = a) \\
&\quad \times \Pr(L_{k-1} = l_{k-1} | Y_{k-1}(\overline{a}) = 0, \overline{L}_{k-2} = \bar{l}_{k-2}, \forall j < k-1 : A_j = a) \\
&\quad \times \prod_{j=0}^{k-2} \Pr(Y_{j+1}(\overline{a}) = 1 | Y_j(\overline{a}) = 0, \overline{L}_j = \bar{l}_j, \forall i < j : A_i = a) \\
&\quad \times \Pr(L_j = l_j | Y_j(\overline{a}) = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j : A_i = a) \\
&\ \ \vdots
\end{aligned}
$$

(by repeating previous three steps, under sequential conditional exchangeability)

$$= \Pr(Y_{k+1}(\overline{a}) = 1, Y_k(\overline{a}) = 0)$$

and, similarly,

$$
\begin{aligned}
&\sum_{\bar{l}_k} \frac{\Pr(Y_{k+1} = 1, Y_k = 0, \forall j \leq k : A_j = a', \overline{L}_k = \bar{l}_k)}{\prod_{j=0}^{k} \Pr(A_j = a' | Y_j = 0, \overline{L}_k = \bar{l}_k, \forall i < j : A_i = a')} \\
&= \Pr(Y_{k+1}(\overline{a}') = 1, Y_k(\overline{a}') = 0).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\frac{\mathbb{E}\big[\sum_{k=0}^{K-1} I(A_k = a) W_k I(Y_{k+1} = 1, Y_k = 0, \forall j \leq k : A_j = A_0)\big]}{\mathbb{E}\big[\sum_{k=0}^{K-1} I(A_k = a') W_k I(Y_{k+1} = 1, Y_k = 0, \forall j \leq k : A_j = A_0)\big]} \\
&\quad = \frac{\sum_{k=0}^{K-1} \Pr(Y_{k+1}(\overline{a}) = 1, Y_k(\overline{a}) = 0)}{\sum_{k=0}^{K-1} \Pr(Y_{k+1}(\overline{a}') = 1, Y_k(\overline{a}') = 0)} \\
&\quad = \frac{\sum_{k=0}^{K-1} \Pr(Y_{k+1}(\overline{a}) = 1 | Y_k(\overline{a}) = 0) \prod_{j=1}^{k} \Pr(Y_j(\overline{a}) = 0 | Y_{j-1}(\overline{a}) = 0)}{\sum_{k=0}^{K-1} \Pr(Y_{k+1}(\overline{a}') = 1 | Y_k(\overline{a}') = 0) \prod_{j=1}^{k} \Pr(Y_j(\overline{a}') = 0 | Y_{j-1}(\overline{a}') = 0)} \\
&\quad = \frac{\sum_{k=0}^{K-1} \theta_a (1 - \theta_a)^k}{\sum_{k=0}^{K-1} \theta_{a'} (1 - \theta_{a'})^k} \qquad \text{(H4)} \\
&\quad = \frac{1 - (1 - \theta_a)^K}{1 - (1 - \theta_{a'})^K} \quad \text{(since } (1 - r) \textstyle\sum_{k=l}^{u} ar^k = a(r^l - r^{u+1}) \text{ for any real } a, r\text{)}
\end{aligned}
$$

For the denominator, we have

$$
\begin{aligned}
&\frac{\mathbb{E}\big[I(A_0 = a)\sum_{k=0}^{K-1} W_k S_k | \forall j : Y_j = 0 \Rightarrow A_j = A_0\big]}{\mathbb{E}\big[I(A_0 = a')\sum_{k=0}^{K-1} W_k S_k | \forall j : Y_j = 0 \Rightarrow A_j = A_0\big]} \\
&\quad = \frac{\mathbb{E}\big[\sum_{k=0}^{K-1} I(A_k = a) W_k S_k | \forall j : Y_j = 0 \Rightarrow A_j = A_0\big]}{\mathbb{E}\big[\sum_{k=0}^{K-1} I(A_k = a') W_k S_k | \forall j : Y_j = 0 \Rightarrow A_j = A_0\big]} \\
&\quad = \frac{\sum_{k=0}^{K-1} \mathbb{E}\big[I(A_k = a) W_k S_k | \forall j : Y_j = 0 \Rightarrow A_j = A_0\big]}{\sum_{k=0}^{K-1} \mathbb{E}\big[I(A_k = a') W_k S_k | \forall j : Y_j = 0 \Rightarrow A_j = A_0\big]} \\
&\quad = \frac{\begin{array}{l}\sum_{k=0}^{K-1} \mathbb{E}\big[I(A_k = a) W_k S_k | Y_k = 0, \forall j \leq k : A_j = A_0\big] \\ \quad \times \Pr(Y_k = 0 | \forall j : Y_j = 0 \Rightarrow A_j = A_0)\end{array}}{\begin{array}{l}\sum_{k=0}^{K-1} \mathbb{E}\big[I(A_k = a') W_k S_k | Y_k = 0, \forall j \leq k : A_j = A_0\big] \\ \quad \times \Pr(Y_k = 0 | \forall j : Y_j = 0 \Rightarrow A_j = A_0)\end{array}} && \text{(by S4)} \\
&\quad = \frac{\begin{array}{l}\sum_{k=0}^{K-1} \mathbb{E}\big[I(A_k = a) W_k S_k | Y_k = 0, \forall j \leq k : A_j = A_0\big] \\ \quad \times \Pr(Y_k = 0, \forall j \leq k : A_j = A_0)\end{array}}{\begin{array}{l}\sum_{k=0}^{K-1} \mathbb{E}\big[I(A_k = a') W_k S_k | Y_k = 0, \forall j \leq k : A_j = A_0\big] \\ \quad \Pr(Y_k = 0, \forall j \leq k : A_j = A_0)\end{array}} \\
&\quad = \frac{\sum_{k=0}^{K-1} \mathbb{E}\big[W_k S_k | Y_k = 0, \forall j \leq k : A_j = a\big] \Pr(Y_k = 0, \forall j \leq k : A_j = a)}{\sum_{k=0}^{K-1} \mathbb{E}\big[W_k S_k | Y_k = 0, \forall j \leq k : A_j = a'\big] \Pr(Y_k = 0, \forall j \leq k : A_j = a')} \\
&\quad = \frac{\displaystyle\sum_{k=0}^{K-1} \sum_{\bar{l}_k} \begin{array}{c}\mathbb{E}\big[S_k | Y_k = 0, \overline{L}_k = \bar{l}_k, \forall j \leq k : A_j = a\big] \\ \times \dfrac{\Pr(Y_k = 0, \overline{L}_k = \bar{l}_k, \forall j \leq k : A_j = a)}{\prod_{j=0}^{k} \Pr(A_j = a | Y_j = 0, \overline{L}_j = \bar{l}_j, \forall i < j : A_i = a)}\end{array}}{\displaystyle\sum_{k=0}^{K-1} \sum_{\bar{l}_k} \begin{array}{c}\mathbb{E}\big[S_k | Y_k = 0, \overline{L}_k = \bar{l}_k, \forall j \leq k : A_j = a'\big] \\ \times \dfrac{\Pr(Y_k = 0, \overline{L}_k = \bar{l}_k, \forall j \leq k : A_j = a')}{\prod_{j=0}^{k} \Pr(A_j = a' | Y_j = 0, \overline{L}_j = \bar{l}_j, \forall i < j : A_i = a')}\end{array}} \\
&\quad = \frac{\displaystyle\sum_{k=0}^{K-1} \sum_{\bar{l}_k} \delta \frac{\Pr(Y_k = 0, \overline{L}_k = \bar{l}_k, \forall j \leq k : A_j = a)}{\prod_{j=0}^{k} \Pr(A_j = a | Y_j = 0, \overline{L}_j = \bar{l}_j, \forall i < j : A_i = a)}}{\displaystyle\sum_{k=0}^{K-1} \sum_{\bar{l}_k} \delta \frac{\Pr(Y_k = 0, \overline{L}_k = \bar{l}_k, \forall j \leq k : A_j = a')}{\prod_{j=0}^{k} \Pr(A_j = a' | Y_j = 0, \overline{L}_j = \bar{l}_j, \forall i < j : A_i = a')}} && \text{(by S4)} \\
&\quad = \frac{\sum_{k=0}^{K-1} \sum_{\bar{l}_k} \delta \prod_{j=0}^{k} \begin{array}{l}\Pr(Y_j = 0, L_j = l_j | Y_{j-1} = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \\ \quad \forall i < j : A_i = a)\end{array}}{\sum_{k=0}^{K-1} \sum_{\bar{l}_k} \delta \prod_{j=0}^{k} \begin{array}{l}\Pr(Y_j = 0, L_j = l_j | Y_{j-1} = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \\ \quad \forall i < j : A_i = a')\end{array}}
\end{aligned}
$$

$$= \frac{\displaystyle\sum_{k=0}^{K-1}\sum_{\bar{l}_k}\delta\prod_{j=0}^{k} \begin{array}{c}\Pr(L_j = l_j|Y_j = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j : A_i = a) \times \\ \Pr(Y_j = 0|Y_{j-1} = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j : A_i = a)\end{array}}{\displaystyle\sum_{k=0}^{K} -1\sum_{\bar{l}_k}\delta\prod_{j=0}^{k} \begin{array}{c}\Pr(L_j = l_j|Y_j = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j : A_i = a') \times \\ \Pr(Y_j = 0|Y_{j-1} = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j : A_i = a')\end{array}}$$

$$= \frac{\displaystyle\sum_{k=0}^{K-1}\sum_{\bar{l}_k}\delta\prod_{j=0}^{k} \begin{array}{c}\Pr(L_j = l_j|Y_j(\overline{a}) = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j : A_i = a) \times \\ \Pr(Y_j(\overline{a}) = 0|Y_{j-1}(\overline{a}) = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j : A_i = a)\end{array}}{\displaystyle\sum_{k=0}^{K-1}\sum_{\bar{l}_k}\delta\prod_{j=0}^{k} \begin{array}{c}\Pr(L_j = l_j|Y_j(\overline{a}') = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j : A_i = a') \times \\ \Pr(Y_j(\overline{a}') = 0|Y_{j-1}(\overline{a}') = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j : A_i = a')\end{array}}$$

(by consistency)

$$= \frac{\displaystyle\sum_{k=0}^{K-1}\sum_{\bar{l}_{k-1}}\delta\prod_{j=0}^{k} \begin{array}{c}\Pr(Y_j(\overline{a}) = 0|Y_{j-1}(\overline{a}) = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j : A_i = a) \times \\ \Pr(L_{j-1} = l_{j-1}|Y_{j-1}(\overline{a}) = 0, \overline{L}_{j-2} = \bar{l}_{j-2}, \forall i < j-1 : A_i = a)\end{array}}{\displaystyle\sum_{k=0}^{K-1}\sum_{\bar{l}_{k-1}}\delta\prod_{j=0}^{k} \begin{array}{c}\Pr(Y_j(\overline{a}') = 0|Y_{j-1}(\overline{a}') = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j : A_i = a') \times \\ \Pr(L_{j-1} = l_{j-1}|Y_{j-1}(\overline{a}') = 0, \overline{L}_{j-2} = \bar{l}_{j-2}, \forall i < j-1 : A_i = a')\end{array}}$$

$$= \frac{\displaystyle\sum_{k=0}^{K-1}\sum_{\bar{l}_{k-1}}\delta\prod_{j=0}^{k} \begin{array}{c}\Pr(Y_j(\overline{a}) = 0|Y_{j-1}(\overline{a}) = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j-1 : A_i = a) \times \\ \Pr(L_{j-1} = l_{j-1}|Y_{j-1}(\overline{a}) = 0, \overline{L}_{j-2} = \bar{l}_{j-2}, \forall i < j-1 : A_i = a)\end{array}}{\displaystyle\sum_{k=0}^{K-1}\sum_{\bar{l}_{k-1}}\delta\prod_{j=0}^{k} \begin{array}{c}\Pr(Y_j(\overline{a}') = 0|Y_{j-1}(\overline{a}') = 0, \overline{L}_{j-1} = \bar{l}_{j-1}, \forall i < j-1 : A_i = a') \times \\ \Pr(L_{j-1} = l_{j-1}|Y_{j-1}(\overline{a}') = 0, \overline{L}_{j-2} = \bar{l}_{j-2}, \forall i < j-1 : A_i = a')\end{array}}$$

(by sequential conditional exchangeability)

$$= \frac{\sum_{k=0}^{K-1}\sum_{\bar{l}_{k-1}}\delta\prod_{j=0}^{k} \begin{array}{c}\Pr(Y_j(\overline{a}) = 0, L_{j-1} = l_{j-1}|Y_{j-1}(\overline{a}) = 0, \overline{L}_{j-2} = \bar{l}_{j-2}, \\ \forall i < j-1 : A_i = a)\end{array}}{\sum_{k=0}^{K-1}\sum_{\bar{l}_{k-1}}\delta\prod_{j=0}^{k} \begin{array}{c}\Pr(Y_j(\overline{a}') = 0, L_{j-1} = l_{j-1}|Y_{j-1}(\overline{a}') = 0, \overline{L}_{j-2} = \bar{l}_{j-2}, \\ \forall i < j-1 : A_i = a')\end{array}}$$

$$= \frac{\displaystyle\sum_{k=0}^{K-1}\sum_{\bar{l}_{k-2}}\delta \begin{array}{c}\Pr(Y_k(\overline{a}) = 0|Y_{k-1}(\overline{a}) = 0, \overline{L}_{k-2} = \bar{l}_{k-2}, \forall i < k-1 : A_i = a) \times \\ \prod_{j=0}^{k-1}\Pr(Y_j(\overline{a}) = 0, L_{j-1} = l_{j-1}|Y_{j-1}(\overline{a}) = 0, \overline{L}_{j-2} = \bar{l}_{j-2}, \\ \forall i < j-1 : A_i = a)\end{array}}{\displaystyle\sum_{k=0}^{K-1}\sum_{\bar{l}_{k-2}}\delta \begin{array}{c}\Pr(Y_k(\overline{a}') = 0|Y_{k-1}(\overline{a}') = 0, \overline{L}_{k-2} = \bar{l}_{k-2}, \forall i < k-1 : A_i = a') \times \\ \prod_{j=0}^{k-1}\Pr(Y_j(\overline{a}') = 0, L_{j-1} = l_{j-1}|Y_{j-1}(\overline{a}') = 0, \overline{L}_{j-2} = \bar{l}_{j-2}, \\ \forall i < j-1 : A_i = a')\end{array}}$$

$$\vdots$$

(by sequential conditional exchangeability)

$$
\begin{aligned}
&= \frac{\sum_{k=0}^{K-1} \delta \Pr(Y_k(\overline{a}) = 0)}{\sum_{k=0}^{K-1} \delta \Pr(Y_k(\overline{a}') = 0)} \\
&= \frac{\sum_{k=0}^{K-1} \Pr(Y_k(\overline{a}) = 0)}{\sum_{k=0}^{K-1} \Pr(Y_k(\overline{a}') = 0)} \\
&= \frac{1 + \sum_{k=1}^{K-1} \prod_{j=1}^{k} \Pr(Y_j(\overline{a}) = 0 | Y_{j-1}(\overline{a}) = 0)}{1 + \sum_{k=1}^{K-1} \prod_{j=1}^{k} \Pr(Y_j(\overline{a}') = 0 | Y_{j-1}(\overline{a}') = 0)} \\
&= \frac{1 + \sum_{k=1}^{K-1} (1 - \theta_a)^k}{1 + \sum_{k=1}^{K} (1 - \theta_{a'})^k} && \text{(by H4)} \\
&= \frac{1 + [1 - \theta_a - (1 - \theta_a)^{K-1}]/\theta_a}{1 + [1 - \theta_{a'} - (1 - \theta_{a'})^{K-1}]/\theta_{a'}} \\
& \qquad\qquad \text{(since } (1 - r) \textstyle\sum_{k=l}^{u} ar^k = a(r^l - r^{u+1}) \text{ for any real } a, r) \\
&= \frac{\theta_{a'}(1 - (1 - \theta_a)^{K-1})}{\theta_a(1 - (1 - \theta_{a'})^{K-1})}.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\frac{\dfrac{\mathbb{E}[\sum_{k=0}^{K-1} I(A_k = a) W_k I(Y_{k+1} = 1, Y_k = 0, \forall j \leq k : A_j = A_0) | Y_K = 1]}{\mathbb{E}[\sum_{k=0}^{K-1} I(A_k = 1 - a) W_k I(Y_{k+1} = 1, Y_k = 0, \forall j \leq k : A_j = A_0) | Y_K = 1]}}{\dfrac{\mathbb{E}[\sum_{k=0}^{K-1} I(A_k = a) W_k S_k]}{\mathbb{E}[\sum_{k=0}^{K-1} I(A_k = 1 - a) W_k S_k]}} \\
&\quad = \frac{1 - (1 - \theta_a)^{K-1}}{1 - (1 - \theta_{a'})^{K-1}} \times \frac{\theta_a(1 - (1 - \theta_{a'})^{K-1})}{\theta_{a'}(1 - (1 - \theta_a)^{K-1})} \\
&\quad = \theta_a / \theta_{a'},
\end{aligned}
$$

which completes the proof. □

## S10.3 Identification results for exact 1:$M$ matching strategies

*Intention-to-treat effect*

In this subsection, cases are defined by $Y_K = 1$ and have baseline exposure $A_0$. All cases are assigned a (possibly variable) number $M \geq 0$ of control exposures $A_i'$, $i = 1, ..., M$, subject to

$$
\left.
\begin{gathered}
\Pr(M > 0 | Y_K = 1) > 0 \text{ and} \\
M \perp\!\!\!\perp A_0 | (L_0, Y_K = 1) \text{ and} \\
\forall l, a, a' : \Pr(A_i' = a' | L_0 = l, A_0 = a, Y_K = 1, M, M > 0) \\
= \Pr(A_0 = a' | L_0 = l),
\end{gathered}
\right\} \qquad \text{(M1)}
$$

or

$$\left.\begin{array}{c}\Pr(M > 0|Y_K = 1) > 0 \text{ and}\\ M \perp\!\!\!\perp A_0|(L_0, Y_K = 1) \text{ and}\\ \forall l, a, a' : \Pr(A'_i = a'|L_0 = l, A_0 = a, Y_K = 1, M, M > 0)\\ = \Pr(A_0 = a'|L_0 = l, Y_K = 0),\end{array}\right\} \tag{M2}$$

or

$$\left.\begin{array}{c}\Pr(M > 0|Y_K = 1) > 0 \text{ and}\\ M \perp\!\!\!\perp A_0|(L_0, Y_K = 1, J) \text{ and}\\ \forall l, a, a' : \Pr(A'_i = a'|L_0 = l, A_0 = a, \overline{Y}_K, J = j, M, M > 0)\\ = \Pr(A_0 = a'|L_0 = l, Y_j = 0),\\ \text{where } J = \max\{k = 0, 1, ..., K : Y_k = 0\}.\end{array}\right\} \tag{M3}$$

That is, cases are matched with subjects that have the same baseline covariate level and who are alive at baseline (M1), at the end of study (M2), or whenever the case is alive (M3).

For simplicity, it is assumed below that the variables are discrete. The results can however be extended to more general distributions.

**Theorem 10.7** (Case-base sampling for marginal intention-to-treat effect)**.** *If M1 and BCE hold and*

$$\frac{\Pr(Y_K = 1|L_0 = l, A_0 = 1)}{\Pr(Y_K = 1|L_0 = l, A_0 = 0)} = \theta \tag{H1}$$

*for all $l$ and some constant $\theta$, then*

$$\frac{\mathbb{E}\big[\sum_{i=1}^{M} I(A'_i = 0, A_0 = 1)\big|Y_K = 1, M > 0\big]}{\mathbb{E}\big[\sum_{i=1}^{M} I(A'_i = 1, A_0 = 0)\big|Y_K = 1, M > 0\big]} = \frac{\Pr(Y_K(1) = 1)}{\Pr(Y_K(0) = 1)}.$$

*Proof.* We have

$$\begin{aligned}&\frac{\mathbb{E}\big[\sum_{i=1}^{M} I(A'_i = 0, A_0 = 1)\big|Y_K = 1, M > 0\big]}{\mathbb{E}\big[\sum_{i=1}^{M} I(A'_i = 1, A_0 = 0)\big|Y_K = 1, M > 0\big]}\\ &\qquad = \frac{\mathbb{E}\big[\sum_{i=1}^{M} I(A'_i = 0)\big|A_0 = 1, Y_K = 1, M > 0\big]}{\mathbb{E}\big[\sum_{i=1}^{M} I(A'_i = 1)\big|A_0 = 0, Y_K = 1, M > 0\big]}\\ &\qquad\quad \times \text{Odds}(A_0 = 1|Y_K = 1, M > 0),\end{aligned}$$

where

$$\frac{\mathbb{E}\big[\sum_{i=1}^{M} I(A'_i = 0)\big|A_0 = 1, Y_K = 1, M > 0\big]}{\mathbb{E}\big[\sum_{i=1}^{M} I(A'_i = 1)\big|A_0 = 0, Y_K = 1, M > 0\big]}$$

$$= \frac{\begin{array}{l}\mathbb{E}\big[\sum_{i=1}^{m} I(A_i'=0)|A_0=1,Y_K=1,M=m\big] \\ \quad \times \Pr(M=m|A_0=1,Y_K=1,M>0)\end{array}}{\sum_{m>0} \begin{array}{l}\mathbb{E}\big[\sum_{i=1}^{m} I(A_i'=1)|A_0=0,Y_K=1,M=m\big] \\ \quad \times \Pr(M=m|A_0=0,Y_K=1,M>0)\end{array}}$$

$$= \frac{\sum_{m>0}\sum_{i=1}^{m}\sum_{l} \begin{array}{l}\Pr(A_i'=0|L_0=l,A_0=1,Y_K=1,M=m) \\ \quad \times \Pr(M=m,L_0=l|A_0=1,Y_K=1,M>0)\end{array}}{\sum_{m>0}\sum_{i=1}^{m}\sum_{l} \begin{array}{l}\Pr(A_i'=1|L_0=l,A_0=0,Y_K=1,M=m) \\ \quad \times \Pr(M=m,L_0=l|A_0=0,Y_K=1,M>0)\end{array}}$$

$$= \frac{\sum_{m>0}\sum_{i=1}^{m}\sum_{l} \begin{array}{l}\Pr(A_0=0|L_0=l) \\ \quad \times \Pr(M=m,L_0=l|A_0=1,Y_K=1,M>0)\end{array}}{\sum_{m>0}\sum_{i=1}^{m}\sum_{l} \begin{array}{l}\Pr(A_0=1|L_0=l) \\ \quad \times \Pr(M=m,L_0=l|A_0=0,Y_K=1,M>0)\end{array}} \qquad \text{(by M1)}$$

$$= \frac{\sum_{m>0}\sum_{i=1}^{m}\sum_{l} \begin{array}{l}\Pr(A_0=0|L_0=l) \\ \quad \times \Pr(M=m,L_0=l,A_0=1|Y_K=1)\end{array}}{\sum_{m>0}\sum_{i=1}^{m}\sum_{l} \begin{array}{l}\Pr(A_0=1|L_0=l) \\ \quad \times \Pr(M=m,L_0=l,A_0=0|Y_K=1)\end{array}} \times \frac{1}{\mathrm{Odds}(A_0=1|Y_K=1,M>0)}$$

$$= \frac{\sum_{m>0}\sum_{i=1}^{m}\sum_{l} q(l,m)\Pr(Y_K=1|L_0=l,A_0=1)}{\sum_{m>0}\sum_{i=1}^{m}\sum_{l} q(l,m)\Pr(Y_K=1|L_0=l,A_0=0)} \times \frac{1}{\mathrm{Odds}(A_0=1|Y_K=1,M>0)}$$

(under M1 and definition of $q(l,m)$ (see below))

$$= \frac{\sum_{m>0}\sum_{i=1}^{m}\sum_{l} q(l,m)\theta\Pr(Y_K=1|L_0=l,A_0=0)}{\sum_{m>0}\sum_{i=1}^{m}\sum_{l} q(l,m)\Pr(Y_K=1|L_0=l,A_0=0)} \times \frac{1}{\mathrm{Odds}(A_0=1|Y_K=1,M>0)} \qquad \text{(by H1)}$$

$$= \frac{\theta}{\mathrm{Odds}(A_0=1|Y_K=1,M>0)}$$

where $q(l,m) = \Pr(M=m|L_0=l,Y_K=1)\Pr(A_0=0|L_0=l)\Pr(A_0=1|L_0=l)\Pr(L_0=l)$.

It follows that

$$\frac{\mathbb{E}\big[\sum_{i=1}^{M} I(A_i'=0,A_0=1)|Y_K=1,M>0\big]}{\mathbb{E}\big[\sum_{i=1}^{M} I(A_i'=1,A_0=0)|Y_K=1,M>0\big]} = \frac{\Pr(Y_K=1|L_0,A_0=1)}{\Pr(Y_K=1|L_0,A_0=0)}$$

$$
\begin{aligned}
&= \frac{\Pr(Y_K(1) = 1|L_0, A_0 = 1)}{\Pr(Y_K(0) = 1|L_0, A_0 = 0)} \\
&\qquad\qquad \text{(by consistency)} \\
&= \frac{\Pr(Y_K(1) = 1|L_0)}{\Pr(Y_K(0) = 1|L_0)} \\
&\qquad \text{(by baseline conditional exchangeability)} \\
&= \frac{\Pr(Y_K(1) = 1)}{\Pr(Y_K(0) = 1)}.
\end{aligned}
$$

□

**Theorem 10.8** (Survivor sampling for conditional intention-to-treat effect)**.** *Suppose M2 and BCE hold. If*

$$\frac{\text{Odds}(Y_K = 1|L_0, A_0 = 1)}{\text{Odds}(Y_K = 1|L_0, A_0 = 0)} = \theta \tag{H5}$$

*for some constant $\theta$, then*

$$\frac{\mathbb{E}\big[\sum_{i=1}^{M} I(A_i' = 0, A_0 = 1)\big|Y_K = 1, M > 0\big]}{\mathbb{E}\big[\sum_{i=1}^{M} I(A_i' = 1, A_0 = 0)\big|Y_K = 1, M > 0\big]} = \frac{\text{Odds}(Y_K(1) = 1|L_0)}{\text{Odds}(Y_K(0) = 1|L_0)}.$$

*Proof.* We have

$$
\begin{aligned}
&\frac{\mathbb{E}\big[\sum_{i=1}^{M} I(A_i' = 0, A_0 = 1)\big|Y_K = 1, M > 0\big]}{\mathbb{E}\big[\sum_{i=1}^{M} I(A_i' = 1, A_0 = 0)\big|Y_K = 1, M > 0\big]} \\
&\qquad = \frac{\mathbb{E}\big[\sum_{i=1}^{M} I(A_i' = 0)\big|A_0 = 1, Y_K = 1, M > 0\big]}{\mathbb{E}\big[\sum_{i=1}^{M} I(A_i' = 1)\big|A_0 = 0, Y_K = 1, M > 0\big]} \\
&\qquad\qquad \times \text{Odds}(A_0 = 1|Y_K = 1, M > 0),
\end{aligned}
$$

where

$$
\begin{aligned}
&\frac{\mathbb{E}\big[\sum_{i=1}^{M} I(A_i' = 0)\big|A_0 = 1, Y_K = 1, M > 0\big]}{\mathbb{E}\big[\sum_{i=1}^{M} I(A_i' = 1)\big|A_0 = 0, Y_K = 1, M > 0\big]} \\
&\quad = \frac{\displaystyle\sum_{m>0} \begin{aligned}&\mathbb{E}\big[\textstyle\sum_{i=1}^{m} I(A_i' = 0)\big|A_0 = 1, Y_K = 1, M = m\big] \\ &\times \Pr(M = m|A_0 = 1, Y_K = 1)\end{aligned}}{\displaystyle\sum_{m>0} \begin{aligned}&\mathbb{E}\big[\textstyle\sum_{i=1}^{m} I(A_i' = 1)\big|A_0 = 0, Y_K = 1, M = m\big] \\ &\times \Pr(M = m|A_0 = 0, Y_K = 1)\end{aligned}}
\end{aligned}
$$

$$= \frac{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{l} \begin{array}{l}\Pr(A_i' = 0 | L_0 = l, A_0 = 1, Y_K = 1, M = m)\\ \quad \times \Pr(M = m, L_0 = l | A_0 = 1, Y_K = 1, M > 0)\end{array}}{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{l} \begin{array}{l}\Pr(A_i' = 1 | L_0 = l, A_0 = 0, Y_K = 1, M = m)\\ \quad \times \Pr(M = m, L_0 = l | A_0 = 0, Y_K = 1, M > 0)\end{array}}$$

$$= \frac{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{l} \begin{array}{l}\Pr(A_0 = 0 | L_0 = l, Y_K = 0)\\ \quad \times \Pr(M = m, L_0 = l | A_0 = 1, Y_K = 1, M > 0)\end{array}}{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{l} \begin{array}{l}\Pr(A_0 = 1 | L_0 = l, Y_K = 0)\\ \quad \times \Pr(M = m, L_0 = l | A_0 = 0, Y_K = 1, M > 0)\end{array}} \qquad \text{(by M2)}$$

$$= \frac{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{l} \frac{\begin{array}{c}\Pr(Y_K = 0 | L_0 = 0, A_0 = 0)\Pr(A_0 = 0 | L_0 = l)\\ \times \Pr(M = m, L_0 = l, A_0 = 1 | Y_K = 1)\end{array}}{\Pr(Y_K = 0 | L_0 = l)}}{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{l} \frac{\begin{array}{c}\Pr(Y_K = 0 | L_0 = 0, A_0 = 1)\Pr(A_0 = 1 | L_0 = l)\\ \times \Pr(M = m, L_0 = l, A_0 = 0 | Y_K = 1)\end{array}}{\Pr(Y_K = 0 | L_0 = l)}} \times \frac{1}{\text{Odds}(A_0 = 1 | Y_K = 1, M > 0)}$$

$$= \frac{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{l} q(l,m) \begin{array}{l}\Pr(Y_K = 1 | L_0 = l, A_0 = 1)\\ \quad \times \Pr(Y_K = 0 | L_0 = 0, A_0 = 0)\end{array}}{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{l} q(l,m) \begin{array}{l}\Pr(Y_K = 1 | L_0 = l, A_0 = 0)\\ \quad \times \Pr(Y_K = 0 | L_0 = 0, A_0 = 1)\end{array}} \times \frac{1}{\text{Odds}(A_0 = 1 | Y_K = 1, M > 0)}$$

(under M2 and definition of $q(l,m)$ (see below))

$$= \frac{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{l} \begin{array}{l}q(l,m)\theta\Pr(Y_K = 1 | L_0 = l, A_0 = 0)\\ \quad \times \Pr(Y_K = 0 | L_0 = 0, A_0 = 1)\end{array}}{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{l} \begin{array}{l}q(l,m)\Pr(Y_K = 1 | L_0 = l, A_0 = 0)\\ \quad \times \Pr(Y_K = 0 | L_0 = 0, A_0 = 1)\end{array}} \times \frac{1}{\text{Odds}(A_0 = 1 | Y_K = 1, M > 0)} \qquad \text{(by H5)}$$

$$= \frac{\theta}{\text{Odds}(A_0 = 1 | Y_K = 1, M > 0)}$$

where $q(l,m) = \Pr(M = m | L_0 = l, Y_K = 1)\Pr(A_0 = 0 | L_0 = l)\Pr(A_0 = 1 | L_0 = l)\Pr(L_0 = l)/\Pr(Y_K = 0 | L_0 = l)$.

From the definition of $\theta$, it follows that

$$\begin{aligned}\frac{\mathbb{E}[\sum_{i=1}^{M} I(A_i'=0,A_0=1)|Y_K=1,M>0]}{\mathbb{E}[\sum_{i=1}^{M} I(A_i'=1,A_0=0)|Y_K=1,M>0]} &= \frac{\text{Odds}(Y_K(1)=1|L_0,A_0=1)}{\text{Odds}(Y_K(0)=1|L_0,A_0=0)} \\ &\qquad\text{(by consistency)} \\ &= \frac{\text{Odds}(Y_K(1)=1|L_0)}{\text{Odds}(Y_K(0)=1|L_0)} \\ &\qquad\text{(by baseline conditional exchangeability)} \\ &= \frac{\text{Odds}(Y_K(1)=1)}{\text{Odds}(Y_K(0)=1)}.\end{aligned}$$

□

**Theorem 10.9** (Risk-set sampling for conditional intention-to-treat effect). *Suppose M3 and BCE hold. If*

$$\frac{\Pr(Y_{j+1}=1|L_0,A_0=1,Y_j=0)}{\Pr(Y_{j+1}=1|L_0,A_0=0,Y_j=0)}=\theta \tag{H6}$$

*for* $j=0,1,...,K$ *and some constant* $\theta$*, then*

$$\frac{\mathbb{E}[\sum_{i=1}^{M} I(A_i'=0,A_0=1)|Y_K=1,M>0]}{\mathbb{E}[\sum_{i=1}^{M} I(A_i'=1,A_0=0)|Y_K=1,M>0]} = \frac{\Pr(Y_{j+1}(1)=1|L_0,Y_j(1)=0)}{\Pr(Y_{j+1}(0)=1|L_0,Y_j(0)=0)}.$$

*Proof.* If $J=\max\{k=0,1,...,K:Y_k=0\}$, then

$$\begin{aligned}&\frac{\mathbb{E}[\sum_{i=1}^{M} I(A_i'=0,A_0=1)|Y_K=1,M>0]}{\mathbb{E}[\sum_{i=1}^{M} I(A_i'=1,A_0=0)|Y_K=1,M>0]} \\ &\quad= \frac{\sum_{m>0} \mathbb{E}[\sum_{i=1}^{m} I(A_i'=0,A_0=1)|Y_K=1,M=m] \times \Pr(M=m|Y_K=1,M>0)}{\sum_{m>0} \mathbb{E}[\sum_{i=1}^{m} I(A_i'=1,A_0=0)|Y_K=1,M=m] \times \Pr(M=m|Y_K=1,M>0)} \\ &\quad= \frac{\sum_{m>0} \mathbb{E}[\sum_{i=1}^{m} I(A_i'=0,A_0=1)|Y_K=1,M=m] \times \Pr(M=m,Y_K=1)}{\sum_{m>0} \mathbb{E}[\sum_{i=1}^{m} I(A_i'=1,A_0=0)|Y_K=1,M=m] \times \Pr(M=m,Y_K=1)}\end{aligned}$$

$$= \frac{\displaystyle\sum_{m>0}\sum_{j=0}^{K-1}\sum_{l} \begin{array}{l}\mathbb{E}\big[\sum_{i=1}^{m} I(A'_i=0, A_0=1)\big|L_0=l, J=j, M=m\big] \\ \Pr(L_0=l, J=j, M=m)\end{array}}{\displaystyle\sum_{m>0}\sum_{j=0}^{K-1}\sum_{l} \begin{array}{l}\mathbb{E}\big[\sum_{i=1}^{m} I(A'_i=1, A_0=0)\big|L_0=l, J=j, M=m\big] \\ \quad\times\Pr(L_0=l, J=j, M=m)\end{array}}$$

$$= \frac{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{j=0}^{K-1}\sum_{l} \begin{array}{l}\mathbb{E}\big[I(A'_i=0, A_0=1)\big|L_0=l, J=j, M=m\big] \\ \quad\times\Pr(L_0=l, J=j, M=m)\end{array}}{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{j=0}^{K-1}\sum_{l} \begin{array}{l}\mathbb{E}\big[I(A'_i=1, A_0=0)\big|L_0=l, J=j, M=m\big] \\ \quad\times\Pr(L_0=l, J=j, M=m)\end{array}}$$

$$= \frac{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{j=0}^{K-1}\sum_{l} \begin{array}{l}\mathbb{E}\big[I(A'_i=0, A_0=1)\big|L_0=l, Y_j=0, Y_{j+1}=1, M=m\big] \\ \qquad\times\Pr(L_0=l, Y_j=0, Y_{j+1}=1, M=m)\end{array}}{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{j=0}^{K-1}\sum_{l} \begin{array}{l}\mathbb{E}\big[I(A'_i=1, A_0=0)\big|L_0=l, Y_j=0, Y_{j+1}=1, M=m\big] \\ \qquad\times\Pr(L_0=l, Y_j=0, Y_{j+1}=1, M=m)\end{array}}$$

$$= \frac{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{j=0}^{K-1}\sum_{l} \begin{array}{l}\Pr(A'_i=0|L_0=l, A_0=1, Y_j=0, Y_{j+1}=1, M=m) \\ \qquad\times\Pr(L_0=l, A_0=1, Y_j=0, Y_{j+1}=1, M=m)\end{array}}{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{j=0}^{K-1}\sum_{l} \begin{array}{l}\Pr(A'_i=1|L_0=l, A_0=0, Y_j=0, Y_{j+1}=1, M=m) \\ \qquad\times\Pr(L_0=l, A_0=0, Y_j=0, Y_{j+1}=1, M=m)\end{array}}$$

$$= \frac{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{j=0}^{K-1}\sum_{l} \begin{array}{l}\Pr(A_0=0|L_0=l, Y_j=0) \\ \quad\times\Pr(L_0=l, A_0=1, Y_j=0, Y_{j+1}=1, M=m)\end{array}}{\displaystyle\sum_{m>0}\sum_{i=1}^{m}\sum_{j=0}^{K-1}\sum_{l} \begin{array}{l}\Pr(A_0=1|L_0=l, Y_j=0) \\ \quad\times\Pr(L_0=l, A_0=0, Y_j=0, Y_{j+1}=1, M=m)\end{array}}$$

(by M3)

$$= \frac{\sum_{m>0}\sum_{i=1}^{m}\sum_{j=0}^{K-1}\sum_{l} q_j(l,m)\Pr(Y_{j+1}=1|L_0=l, A_0=1, Y_j=0)}{\sum_{m>0}\sum_{i=1}^{m}\sum_{j=0}^{K-1}\sum_{l} q_j(l,m)\Pr(Y_{j+1}=1|L_0=l, A_0=0, Y_j=0)}$$

(under M3 and definition of $q_j(l,m)$ (see below))

$$= \theta\frac{\sum_{m>0}\sum_{i=1}^{m}\sum_{j=0}^{K-1}\sum_{l} q_j(l,m)\Pr(Y_{j+1}=1|L_0=l, A_0=0, Y_j=0)}{\sum_{m>0}\sum_{i=1}^{m}\sum_{j=0}^{K-1}\sum_{l} q_j(l,m)\Pr(Y_{j+1}=1|L_0=l, A_0=0, Y_j=0)}$$

(by H6)

$$= \theta.$$

where $q_j(l,m) = \Pr(M=m|L_0=l, Y_j=0)\Pr(A_0=1|L_0=l, Y_j=0)\Pr(A_0=$

$0 | L_0 = l, Y_j = 0) \Pr(L_0 = l, Y_j = 0)$.

Thus,

$$\begin{aligned}
&\frac{\mathbb{E}[\sum_{i=1}^{M} I(A'_i = 0, A_0 = 1) | Y_K = 1]}{\mathbb{E}[\sum_{i=1}^{M} I(A'_i = 1, A_0 = 0) | Y_K = 1]} \\
&\qquad = \frac{\Pr(Y_{j+1} = 1 | L_0, A_0 = 1, Y_j = 0)}{\Pr(Y_{j+1} = 1 | L_0, A_0 = 0, Y_j = 0)} \\
&\qquad = \frac{\Pr(Y_{j+1}(1) = 1 | L_0, A_0 = 1, Y_j(1) = 0)}{\Pr(Y_{j+1}(0) = 1 | L_0, A_0 = 0, Y_j(0) = 0)} && \text{(by consistency)} \\
&\qquad = \frac{\Pr(Y_{j+1}(1) = 1 | L_0, Y_j(1) = 0)}{\Pr(Y_{j+1}(0) = 1 | L_0, Y_j(0) = 0)}. \\
&&& \text{(by baseline conditional exchangeability)}
\end{aligned}$$

□

*Per-protocol effect*

In this subsection, an individual qualifies as a case if and only if $Y_K = 1$ and the subject adheres to the protocol that was assigned at baseline (i.e., $A_k = A_0$ for all $k = 0, 1, ..., K$ if $Y_k = 0$). All cases are assigned a (possibly variable) number $M \geq 0$ control exposures $A'_i$, $i = 1, ..., M$, subject to

$$\left.\begin{array}{l}
\Pr(M > 0 | Y_K = 1, \forall j : (Y_j = 0 \Rightarrow A_j = A_0)) > 0 \text{ and} \\
M \perp\!\!\!\perp A_0 | (J, Y_K = 1, \overline{L}_J = \overline{l}_J, \forall i \leq J : A_i = A_0) \text{ and} \\
\forall \overline{l}, a : \Pr(A'_i = a' | \overline{L}_J = \overline{l}_J, \forall j \leq J : A_j = A_0, A_0 = a, \\
\qquad Y_J = 0, J, M, M > 0) \\
\qquad = \Pr(A_J = a' | \overline{L}_J = \overline{l}_J, \forall j \leq J : A_j = A_0, Y_J = 0), \\
\qquad \text{where } J = \max\{k = 0, 1, ..., K : Y_k = 0\}.
\end{array}\right\} \tag{M4}$$

**Theorem 10.10** (Risk-set sampling for conditional per-protocol effect)**.** *Suppose M4 holds. If*

$$\frac{\Pr(Y_{j+1} = 1 | \overline{L}_j = \overline{l}_j, Y_j = 0, \forall i \leq j : A_i = 1)}{\Pr(Y_{j+1} = 1 | \overline{L}_j = \overline{l}_j, Y_j = 0, \forall i \leq j : A_i = 0)} = \theta \tag{H7}$$

*for all* $j, \overline{l}_j$ *and some constant* $\theta$*, then*

$$\frac{\mathbb{E}\Big[\sum_{i=1}^{M} I(A'_i = 0, A_0 = 1) \Big| Y_K = 1, \forall j : (Y_j = 0 \Rightarrow A_j = A_0), M > 0\Big]}{\mathbb{E}\Big[\sum_{i=1}^{M} I(A'_i = 1, A_0 = 0) \Big| Y_K = 1, \forall j : (Y_j = 0 \Rightarrow A_j = A_0), M > 0\Big]}$$

$$= \frac{\Pr(Y_{j+1}(\overline{1}) = 1|\overline{L}_j = \bar{l}_j, Y_j(\overline{1}) = 0, \forall i \leq j : A_i = 1)}{\Pr(Y_{j+1}(\overline{0}) = 1|\overline{L}_j = \bar{l}_j, Y_j(\overline{0}) = 0, \forall i \leq j : A_i = 0)}.$$

*Proof.* Let $J = \max\{k = 0, 1, ..., K : Y_k = 0\}$. Then, for $a = 0, 1$,

$$
\begin{aligned}
&\mathbb{E}\Bigg[\sum_{i=1}^{M} I(A'_i = 1 - a, A_0 = a)\Bigg| Y_K = 1, \forall j \leq J : A_j = A_0, M > 0\Bigg] \\
&\quad = \sum_{j=0}^{K-1} \sum_{\bar{l}_j} \mathbb{E}\Bigg[\sum_{i=1}^{M} I(A'_i = 1 - a, A_0 = a)\Bigg| \\
&\qquad \overline{L}_j = \bar{l}_j, J = j, Y_K = 1, \forall j \leq J : A_j = A_0, M > 0\Bigg] \\
&\qquad \times \Pr(\overline{L}_j = \bar{l}_j, J = j | Y_K = 1, \forall i \leq J : A_i = A_0, M > 0) \\
&\quad = \sum_{j=0}^{K-1} \sum_{\bar{l}_j} \mathbb{E}\Bigg[\sum_{i=1}^{M} I(A'_i = 1 - a, A_0 = a)\Bigg| \\
&\qquad \overline{L}_j = \bar{l}_j, Y_j = 0, Y_{j+1} = 1, \forall j \leq J : A_j = A_0, M > 0\Bigg] \\
&\qquad \times \Pr(\overline{L}_j = \bar{l}_j, Y_j = 0, Y_{j+1} = 1 | Y_K = 1, \forall i \leq J : A_i = A_0, M > 0) \\
&\quad = \sum_{m>0} \sum_{j=0}^{K-1} \sum_{\bar{l}_j} \mathbb{E}\Bigg[\sum_{i=1}^{m} I(A'_i = 1 - a, A_0 = a)\Bigg| \\
&\qquad \overline{L}_j = \bar{l}_j, Y_j = 0, Y_{j+1} = 1, \forall j \leq J : A_j = A_0, M = m\Bigg] \\
&\qquad \times \Pr(M = m, \overline{L}_j = \bar{l}_j, Y_j = 0, Y_{j+1} = 1 | \forall j \leq J : A_j = A_0, M > 0) \\
&\quad = \sum_{m>0} \sum_{u=1}^{m} \sum_{j=0}^{K-1} \sum_{\bar{l}_j} \mathbb{E}\Bigg[I(A'_u = 1 - a, A_0 = a)\Bigg| \\
&\qquad \overline{L}_j = \bar{l}_j, Y_j = 0, Y_{j+1} = 1, \forall j \leq J : A_j = A_0, M = m\Bigg] \\
&\qquad \times \Pr(M = m, \overline{L}_j = \bar{l}_j, Y_j = 0, Y_{j+1} = 1 | \forall j \leq J : A_j = A_0, M > 0) \\
&\quad = \sum_{m>0} \sum_{u=1}^{m} \sum_{j=0}^{K-1} \sum_{\bar{l}_j} \Pr(A'_u = 1 - a | \overline{L}_j = \bar{l}_j, Y_j = 0, Y_{j+1} = 1, \\
&\qquad \forall j \leq J : A_j = a, M = m)
\end{aligned}
$$

$$
\begin{aligned}
&\quad \times \Pr(M = m, A_0 = a, \overline{L}_j = \bar{l}_j, Y_j = 0, Y_{j+1} = 1 | \\
&\quad \forall j \leq J : A_j = A_0, M > 0) \\
&= \sum_{m>0} \sum_{u=1}^{m} \sum_{j=0}^{K-1} \sum_{\bar{l}_j} \Pr(A_0 = 1 - a | Y_j = 0, \overline{L}_j = \bar{l}_j, \\
&\quad \forall i \leq j : A_i = A_0) \\
&\quad \times \Pr(M = m, A_0 = a, \overline{L}_j = \bar{l}_j, Y_j = 0, Y_{j+1} = 1 | \\
&\quad \forall j \leq J : A_j = A_0, M > 0) \qquad \text{(by M4)} \\
&= \sum_{m>0} \sum_{u=1}^{m} \sum_{j=0}^{K-1} \sum_{\bar{l}_j} \Pr(A_0 = 1 - a | Y_j = 0, \overline{L}_j = \bar{l}_j, \forall i \leq j : A_i = A_0) \\
&\quad \times \Pr(M = m, \overline{L}_j = \bar{l}_j, A_0 = a, Y_{j+1} = 1, Y_j = 0, \forall i \leq j : A_i = A_0) \\
&\quad \times \Pr(Y_K = 1, \forall i : (Y_i = 0 \Rightarrow A_i = A_0), M > 0)^{-1} \\
&= \sum_{m>0} \sum_{u=1}^{m} \sum_{j=0}^{K-1} \sum_{\bar{l}_j} \Pr(Y_{j+1} = 1 | \overline{L}_j = \bar{l}_j, A_0 = a, Y_j = 0, \forall i \leq j : A_i = A_0) \\
&\quad \times q_j(\bar{l}_j, m) \Pr(Y_K = 1, \forall i : (Y_i = 0 \Rightarrow A_i = A_0), M > 0)^{-1}, \\
&\qquad \text{(under M4)}
\end{aligned}
$$

where

$$
\begin{aligned}
q_j(\bar{l}_j, m) = \Pr(M = m | \overline{L}_j = \bar{l}_j, Y_j = 0, Y_{j+1} = 1, \forall i \leq j : A_i = A_0) \\
\times \Pr(A_0 = 1 - a | Y_j = 0, \overline{L}_j = \bar{l}_j, \forall i \leq j : A_i = A_0) \\
\times \Pr(A_0 = a | Y_j = 0, \overline{L}_j = \bar{l}_j, \forall i \leq j : A_i = A_0) \\
\times \Pr(\overline{L}_j = \bar{l}_j, Y_j = 0, \forall i \leq j : A_i = A_0).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\frac{\mathbb{E}\Big[\sum_{i=1}^{M} I(A_i' = 0, A_0 = 1) \Big| Y_K = 1, \forall j : (Y_j = 0 \Rightarrow A_j = A_0), M > 0\Big]}{\mathbb{E}\Big[\sum_{i=1}^{M} I(A_i' = 1, A_0 = 0) \Big| Y_K = 1, \forall j : (Y_j = 0 \Rightarrow A_j = A_0), M > 0\Big]} \\
&= \frac{\sum_{m>0} \sum_{u=1}^{m} \sum_{j=0}^{K-1} \sum_{\bar{l}_j} \begin{matrix} \Pr(Y_{j+1} = 1 | \overline{L}_j = \bar{l}_j, A_0 = 1, Y_j = 0, \forall i \leq j : A_i = A_0) \\ \times q_j(\bar{l}_j, m) \Pr(Y_K = 1, \forall i : (Y_i = 0 \Rightarrow A_i = A_0), M > 0)^{-1} \end{matrix}}{\sum_{m>0} \sum_{u=1}^{m} \sum_{j=0}^{K-1} \sum_{\bar{l}_j} \begin{matrix} \Pr(Y_{j+1} = 1 | \overline{L}_j = \bar{l}_j, A_0 = 0, Y_j = 0, \forall i \leq j : A_i = A_0) \\ \times q_j(\bar{l}_j, m) \Pr(Y_K = 1, \forall i : (Y_i = 0 \Rightarrow A_i = A_0), M > 0)^{-1} \end{matrix}}
\end{aligned}
$$

$$= \frac{\displaystyle\sum_{m>0}\sum_{u=1}^{m}\sum_{j=0}^{K-1}\sum_{\bar{l}_j} \begin{array}{c}\Pr(Y_{j+1}=1|\overline{L}_j=\bar{l}_j, A_0=1, Y_j=0, \forall i\leq j: A_i=A_0)\\ \times\, q_j(\bar{l}_j,m)\end{array}}{\displaystyle\sum_{m>0}\sum_{u=1}^{m}\sum_{j=0}^{K-1}\sum_{\bar{l}_j} \begin{array}{c}\Pr(Y_{j+1}=1|\overline{L}_j=\bar{l}_j, A_0=1, Y_j=0, \forall i\leq j: A_i=A_0)\\ \times\, q_j(\bar{l}_j,m)\end{array}}$$

$$= \theta\frac{\displaystyle\sum_{m>0}\sum_{u=1}^{m}\sum_{j=0}^{K-1}\sum_{\bar{l}_j} \begin{array}{c}\Pr(Y_{j+1}=1|\overline{L}_j=\bar{l}_jY_j=0, \forall i\leq j: A_i=0)\\ \times\, q_j(\bar{l}_j,m)\end{array}}{\displaystyle\sum_{m>0}\sum_{u=1}^{m}\sum_{j=0}^{K-1}\sum_{\bar{l}_j} \begin{array}{c}\Pr(Y_{j+1}=1|\overline{L}_j=\bar{l}_j, Y_j=0, \forall i\leq j: A_i=0)\\ \times\, q_j(\bar{l}_j,m)\end{array}}$$

(by H7)

$$= \theta.$$

The desired results follows by consistency. □

## S10.4 Parametric identification by conditional logistic regression for exact or partial 1:$M$ matching

We now allow for the possibility that cases ($Y_K = 1$) are matched to $M \geq 0$ controls on only part of $L_0$. That part of $L_0$ on which exact matching is done will be denoted $L_0^*$; the other part is denoted $L_0'$, so that $L_0 = (L_0^*, L_0')$. The identification result below rests on the assumption that cases are assigned $M \geq 0$ pairs $(A_i', L_i')$ of baseline exposure and baseline covariate data, $i = 1, ..., M$, subject to

$$\left.\begin{array}{c}\Pr(M>0|Y_K=1)>0 \;\text{ and}\\ M \perp\!\!\!\perp (A_0, L_0)|(L_0^*, Y_K=1) \;\text{ and}\\ \forall l, l', a: \Pr(A_i'=a, L_i'=l'|L_0^*=l, L_0', A_0, Y_K=1, M, M>0)\\ = \Pr(A_0=a, L_0'=l'|L_0^*=l, Y_K=0) \;\text{ and}\\ (L_0', A_0), (L_1', A_1'), ..., (L_M', A_M') \text{ are mutually independent}\\ \text{given } (L_0^*, Y_K=1, M>0).\end{array}\right\} \qquad (\text{M2}^*)$$

It is assumed below that the variables are discrete with finite support for simplicity. The results can however be extended to more general distributions.

**Theorem 10.11** (Conditional logistic regression for conditional intention-to-treat effect)**.** *Suppose BCE and M2* hold. For some known real-valued functions* $f_j$*,*

*$j = 1, ..., p$, assume the following model:*

$$\operatorname{logit} \Pr(Y_K(a) = 1 | L_0) = \alpha + \sum_{j=1}^{p} f_j(a, L_0^*, L_0')\beta_j \qquad \text{(Outcome Model)}$$

*For $i = 0, ..., M$, let $X_{i,j} = f_j(A_i', L_0^*, L_i') - f_j(A_0, L_0^*, L_0')$, with $A_0' = A_0$, and assume for any $\gamma_1, ..., \gamma_p \in \mathbb{R}$, not all zero, that*

$$\Pr\left( \bigvee_{i=1}^{M} \left[ \sum_{j=1}^{p} \gamma_j X_{i,j} \neq 0 \right] \middle| Y_K = 1, M > 0 \right) > 0, \qquad \text{(Linear Independence)}$$

*where $\bigvee$ denotes the logical OR operator (i.e., given any indexed collection $(P_i)_{i \in I}$ of propositions, $\bigvee_{i \in I} P_i$ is the proposition that $P_i$ is true for at least one $i \in I$). Then,*

$$\mathbb{E}\left[ -\log \left( 1 + \sum_{i=1}^{M} \exp\left[ \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j \right] \right)^{-1} \middle| Y_K = 1, M > 0 \right]$$

*is uniquely maximized at $\tilde{\beta} = \beta$.*

*Proof.* We first demonstrate that

$$\mathbb{E}\left[ -\log \left( 1 + \sum_{i=1}^{M} \exp\left[ \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j \right] \right)^{-1} \middle| Y_K = 1, M > 0 \right]$$

has at most one maximum by showing that it is strictly concave as a function of $\tilde{\beta}$. Let $X = (X_1, ..., X_M)$ and $X_i = (X_{i,1}, ..., X_{i,p})$, $i = 1, ..., M$. To show that function $f$,

$$\begin{aligned} f(\beta) &= \mathbb{E}\left[ \log \left( 1 + \sum_{i=1}^{M} \exp\left[ \sum_{j=1}^{p} X_{i,j}\beta_j \right] \right)^{-1} \middle| Y_K = 1, M > 0 \right] \\ &= \sum_{m>0} \sum_{x} \log \left( 1 + \sum_{i=1}^{m} \exp\left[ \sum_{j=1}^{p} x_{i,j}\beta_j \right] \right)^{-1} \\ &\qquad \times \Pr(X = x | Y_K = 1, M = m) \Pr(M = m | Y_K = 1, M > 0), \end{aligned}$$

is convex (and $-f$ concave) it suffices to show that its Hessian is positive semidefinite, i.e., that $\sum_{t=1}^{p} \sum_{u=1}^{p} \beta_k \beta_l H_{k,l}(\beta) \geq 0$ for all $\beta \in \mathbb{R}^p$, where

$$H_{k,l}(\beta) = \frac{\partial}{\partial \beta_l} \frac{\partial}{\partial \beta_k} f(\beta).$$

Positive definiteness of the Hessian, i.e., $\sum_{k=1}^{p}\sum_{l=1}^{p}\beta_k\beta_l H_{k,l}(\beta) > 0$ for all $\beta \in \mathbb{R}^p$ such that $\beta_k \neq 0$ for some $k \in \{1, ..., p\}$, implies strict convexity of $f$ (and $-f$ strictly concave).

Letting $g(X_i, \beta) = \exp\big\{\sum_{j=1}^{p} X_{i,j}\beta_j\big\}$ for $i = 1, ..., M$, we have

$$\begin{aligned}
H_{k,l}(\beta) &= \frac{\partial}{\partial\beta_l}\frac{\partial}{\partial\beta_k} f(\beta) \\
&= \frac{\partial}{\partial\beta_l}\sum_{m>0}\sum_{x}\frac{\sum_{i=1}^{m} x_{i,k} g(x_i, \beta)}{1+\sum_{i=1}^{m} g(x_i,\beta)} \\
&\qquad \times \Pr(X = x|Y_K = 1, M = m)\Pr(M = m|Y_K = 1, M > 0) \\
&= \frac{\partial}{\partial\beta_l}\sum_{m>0}\sum_{x}\frac{\sum_{i=1}^{m} x_{i,k} g(x_i, \beta)}{1+\sum_{i=1}^{m} g(x_i,\beta)} \\
&\qquad \times \Pr(X = x|Y_K = 1, M = m)\Pr(M = m|Y_K = 1, M > 0) \\
&= \sum_{m>0}\sum_{x}\left(1+\sum_{i=1}^{m} g(x_i,\beta)\right)^{-2} \\
&\qquad \times \left[\left(1+\sum_{i=1}^{m} g(x_i,\beta)\right)\left(\sum_{i=1}^{m} X_{i,k}X_{i,l}g(x_i,\beta)\right)\right. \\
&\qquad \left. - \left(\sum_{i=1}^{m} X_{i,k}g(x_i,\beta)\right)\left(\sum_{i=1}^{m} X_{i,l}g(x_i,\beta)\right)\right] \\
&\qquad \times \Pr(X = x|Y_K = 1, M = m)\Pr(M = m|Y_K = 1, M > 0),
\end{aligned}$$

so that, with $v_i = \sqrt{g(x_i,\beta)}$ and $w_i = \sum_{j=1}^{p} x_{i,j}\beta_j\sqrt{g(x_i,\beta)}$,

$$\begin{aligned}
&\sum_{k=1}^{p}\sum_{l=1}^{p}\beta_k\beta_l H_{k,l}(\beta) \\
&\quad = \sum_{m>0}\sum_{x}\frac{\Pr(X = x|Y_K = 1, M = m)\Pr(M = m|Y_K = 1, M > 0)}{\left(1+\sum_{i=1}^{m} g(x_i,\beta)\right)^2} \\
&\qquad \times \left[\sum_{k=1}^{p}\sum_{l=1}^{p}\beta_k\beta_l\left(1+\sum_{i=1}^{m} g(x_i,\beta)\right)\left(\sum_{i=1}^{m} x_{i,k}x_{i,l}g(x_i,\beta)\right)\right. \\
&\qquad \left. - \sum_{k=1}^{p}\sum_{l=1}^{p}\beta_k\beta_l\left(\sum_{i=1}^{m} x_{i,k}g(x_i,\beta)\right)\left(\sum_{i=1}^{m} x_{i,l}g(x_i,\beta)\right)\right] \\
&\quad = \sum_{m>0}\sum_{x}\frac{\Pr(X = x|Y_K = 1, M = m)\Pr(M = m|Y_K = 1, M > 0)}{\left(1+\sum_{i=1}^{m} g(x_i,\beta)\right)^2}
\end{aligned}$$

$$
\begin{aligned}
&\times \left[\left(1+\sum_{i=1}^{m} g(x_i,\beta)\right)\left(\sum_{i=1}^{m} g(x_i,\beta)\left(\sum_{k=1}^{p}\beta_k x_{i,k}\right)\left(\sum_{l=1}^{p}\beta_l x_{i,l}\right)\right)\right.\\
&\left.-\left(\sum_{i=1}^{m}\sum_{k=1}^{p}\beta_k x_{i,k} g(x_i,\beta)\right)\left(\sum_{i=1}^{m}\sum_{l=1}^{p}\beta_l x_{i,l} g(x_i,\beta)\right)\right]\\
={}&\sum_{m>0}\sum_{x}\frac{\Pr(X=x|Y_K=1,M=m)\Pr(M=m|Y_K=1,M>0)}{\left(1+\sum_{i=1}^{m} g(x_i,\beta)\right)^2}\\
&\times \left[\left(1+\sum_{i=1}^{m} g(x_i,\beta)\right)\left(\sum_{i=1}^{m}\left(\sum_{k=1}^{p}\beta_k x_{i,k}\sqrt{g(x_i,\beta)}\right)^2\right)\right.\\
&\left.-\left(\sum_{i=1}^{m}\sum_{k=1}^{p}\beta_k x_{i,k} g(x_i,\beta)\right)^2\right]\\
={}&\sum_{m>0}\sum_{x}\frac{\Pr(X=x|Y_K=1,M=m)\Pr(M=m|Y_K=1,M>0)}{\left(1+\sum_{i=1}^{m} g(x_i,\beta)\right)^2}\\
&\times \left[\sum_{i=1}^{m}\left(\sum_{k=1}^{p}\beta_k x_{i,k}\sqrt{g(x_i,\beta)}\right)^2+\left(\sum_{i=1}^{m} v_{i,j}^2\right)\left(\sum_{i=1}^{m} w_{i,j}^2\right)\right.\\
&\left.-\left(\sum_{i=1}^{m} v_{i,j}v_{i,j}\right)^2\right]\\
\geq{}&\sum_{m>0}\sum_{x}\frac{\Pr(X=x|Y_K=1,M=m)\Pr(M=m|Y_K=1,M>0)}{\left(1+\sum_{i=1}^{m} g(x_i,\beta)\right)^2}\\
&\times \sum_{i=1}^{m}\left(\sum_{k=1}^{p}\beta_k x_{i,k}\sqrt{g(x_i,\beta)}\right)^2. \quad \text{(by the Cauchy-Schwarz inequality)}
\end{aligned}
$$

Now,

$$
\begin{aligned}
&\sum_{m>0}\sum_{x}\frac{\Pr(X=x|Y_K=1,M=m)\Pr(M=m|Y_K=1,M>0)}{\left(1+\sum_{i=1}^{m} g(x_i,\beta)\right)^2}\\
&\quad\times \sum_{i=1}^{m}\left(\sum_{k=1}^{p}\beta_k x_{i,k}\sqrt{g(x_i,\beta)}\right)^2\\
&\quad=\sum_{m>0}\sum_{x}\frac{\Pr(X=x|Y_K=1,M=m)\Pr(M=m|Y_K=1,M>0)}{\left(1+\sum_{i=1}^{m} g(x_i,\beta)\right)^2}\\
&\quad\times \sum_{i=1}^{m} g(x_i,\beta)\left(\sum_{k=1}^{p}\beta_k x_{i,k}\right)^2
\end{aligned}
$$

$$= \mathbb{E}\Bigg[\Bigg(1+\sum_{i=1}^{M} g(X_i,\beta)\Bigg)^{-2}\sum_{i=1}^{M} g(X_i,\beta)\Bigg(\sum_{k=1}^{p}\beta_k X_{i,k}\Bigg)^2\Bigg| Y_K=1, M>0\Bigg]$$
$$\geq 0$$

with strict inequality under Linear Independence. Thus,

$$\mathbb{E}\Bigg[-\log\Bigg(1+\sum_{i=1}^{M}\exp\Bigg[\sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j\Bigg]\Bigg)^{-1}\Bigg| Y_K=1, M>0\Bigg]$$

has at most one maximum.

It remains to be shown that

$$\mathbb{E}\Bigg[-\log\Bigg(1+\sum_{i=1}^{M}\exp\Bigg[\sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j\Bigg]\Bigg)^{-1}\Bigg| Y_K=1, M>0\Bigg]$$

is maximized at $\tilde{\beta}=\beta$, i.e., $\partial/\partial\tilde{\beta}_k f(\tilde{\beta})=0$ for all $k=1,...,p$ at $\tilde{\beta}=\beta$.

Now,

$$\begin{aligned}\frac{\partial}{\partial\tilde{\beta}_k} f(\tilde{\beta}) &= \mathbb{E}\Bigg[\frac{\sum_{i=1}^{M} X_{i,k} g(X_i,\tilde{\beta})}{1+\sum_{i=1}^{m} g(X_i,\tilde{\beta})}\Bigg| Y_K=1, M>0\Bigg]\\ &= \sum_{l^*}\sum_{m>0}\mathbb{E}\Bigg[\frac{\sum_{i=1}^{m} X_{i,k} g(X_i,\tilde{\beta})}{1+\sum_{i=1}^{m} g(X_i,\tilde{\beta})}\Bigg| L_0^*=l^*, Y_K=1, M=m\Bigg]\\ &\qquad \times \Pr(L_0^*=l^*, M=m|, Y_K=1, M>0),\end{aligned}$$

where

$$\begin{aligned}&\mathbb{E}\Bigg[\frac{\sum_{i=1}^{m} X_{i,k} g(X_i,\tilde{\beta})}{1+\sum_{i=1}^{m} g(X_i,\tilde{\beta})}\Bigg| L_0^*=l^*, Y_K=1, M=m\Bigg]\\ &\quad = \sum_{l_0,\dots,l_m}\sum_{a_0,\dots,a_m}\frac{\begin{gathered}\sum_{i=1}^{m}[f_k(a_i,l^*,l_i)-f_k(a_0,l^*,l_0)]\\ \times\exp\big\{\sum_{k=1}^{p}[f_k(a_i,l^*,l_i)-f_k(a_0,l^*,l_0)]\tilde{\beta}_k\big\}\end{gathered}}{1+\sum_{i=1}^{m}\exp\big\{\sum_{k=1}^{p}[f_k(a_i,l^*,l_i)-f_k(a_0,l^*,l_0)]\tilde{\beta}_k\big\}}\\ &\qquad \times \Pr(A_0=a_0, A_1'=a_1, ..., A_m=a_m, L_0'=l_0, ..., L_m'=l_m|\\ &\qquad L_0^*=l^*, Y_K=1, M=m)\\ &\quad = \sum_{l_0,\dots,l_m}\sum_{a_0,\dots,a_m}\frac{\begin{gathered}\sum_{i=1}^{m}[f_k(a_i,l^*,l_i)-f_k(a_0,l^*,l_0)]\\ \times\exp\big\{\sum_{k=1}^{p}[f_k(a_i,l^*,l_i)-f_k(a_0,l^*,l_0)]\tilde{\beta}_k\big\}\end{gathered}}{1+\sum_{i=1}^{m}\exp\big\{\sum_{k=1}^{p}[f_k(a_i,l^*,l_i)-f_k(a_0,l^*,l_0)]\tilde{\beta}_k\big\}}\end{aligned}$$

$$
\begin{aligned}
&\times h(a_0, ..., a_M, l_0, ..., l_M)\\
&\times \Pr\Big(A_0 = a_0, A_1' = a_1, ..., A_M = a_M, L_0' = l_0, ..., L_m' = l_m\Big|\\
&\bigvee_{\sigma} [(A_0 = a_{\sigma(0)}, L_0' =_{\sigma(0)}, A_1' = a_{\sigma(1)}, L_1' =_{\sigma(1)}, ..., A_m = a_{\sigma(m)}, L_m' =_{\sigma(m)})],\\
&L_0^* = l^*, Y_K = 1, M = m\Big),
\end{aligned}
$$

where permutation $\sigma$ denotes a bijection from $\{0, 1, ..., M\}$ to itself and

$$
\begin{aligned}
&h(a_0, ..., a_M, l_0, ..., l_M)\\
&\quad = \Pr\Big(\bigvee_{\sigma} [(A_0 = a_{\sigma(0)}, L_0' =_{\sigma(0)}, A_1' = a_{\sigma(1)}, L_1' =_{\sigma(1)}, ..., A_m = a_{\sigma(m)},\\
&\qquad L_m' =_{\sigma(m)})]\Big| L_0^* = l^*, Y_K = 1, M = m\Big).
\end{aligned}
$$

Next, let $B_0 = (L_0', A_0)$ and $B_i = (L_i', A_i')$, $i = 1, 2, ..., M$. Let $b_i = (l_i, a_i)$ for $i = 0, ..., M$. We have

$$
\begin{aligned}
&\Pr\Bigg(B_0 = b_0, , ..., B_M = b_M\Bigg|\\
&\quad \bigvee_{\sigma} [(B_0, ..., B_M) = (b_{\sigma(0)}, ..., b_{\sigma(M)})], L_0^*, Y_K = 1, M, M > 0\Bigg)\\
&\quad = \frac{\Pr(B_0 = b_0, ..., B_M = b_M | L_0^*, Y_K = 1, M > 0)}{\Pr\Big(\bigvee_\sigma [B_0 = b_{\sigma(0)}, ..., B_M = a_{\sigma(M)}]\Big| L_0^*, Y_K = 1, M, M > 0\Big)}\\
&\quad \propto \frac{\Pr(B_0 = b_0, ..., B_M = b_M | L_0^*, Y_K = 1, M > 0)}{\sum_\sigma \Pr\Big(B_0 = b_{\sigma(0)}, ..., B_M = a_{\sigma(M)}\Big| L_0^*, Y_K = 1, M, M > 0\Big)}\\
&\quad = \frac{\prod_{i=0}^M \Pr(B_i = b_i | L_0^*, Y_K = 1, M, M > 0)}{\sum_\sigma \prod_{i=0}^M \Pr(B_i = b_{\sigma(i)} | L_0^*, Y_K = 1, M, M > 0)}\\
&\qquad\qquad \text{(by mutual independence of M2*)}\\
&\quad = \frac{\Pr(B_0 = b_0 | L_0^*, Y_K = 1) \prod_{i=1}^M \Pr(B_0 = b_i | L_0^*, Y_K = 0)}{\sum_\sigma \Pr(B_0 = b_{\sigma(0)} | L_0^*, Y_K = 1) \prod_{i=1}^M \Pr(B_0 = b_{\sigma(i)} | L_0^*, Y_K = 0)}\\
&\qquad\qquad \text{(by M2*)}\\
&\quad = \frac{\Pr(Y_K = 1 | B_0 = b_0, L_0^*) \prod_{i=1}^M [1 - \Pr(Y_K = 1 | B_0 = b_i, L_0^*)]}{\sum_\sigma \Pr(Y_K = 1 | B_0 = b_{\sigma(0)}, L_0^*) \prod_{i=1}^M [1 - \Pr(Y_K = 1 | B_0 = b_{\sigma(i)}, L_0^*)]}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\begin{array}{c}\Pr(Y_K = 1 | L_0 = (L_0^*, l_0), A_0 = a_0) \\ \times \prod_{i=1}^{M} [1 - \Pr(Y_K = 1 | L_0 = (L_0^*, l_i), A_0 = a_i)]\end{array}}{\displaystyle\sum_{\sigma} \begin{array}{c}\Pr(Y_K = 1 | L_0 = (L_0^*, l_{\sigma(0)}), A_0 = a_{\sigma(0)}) \\ \times \prod_{i=1}^{M} [1 - \Pr(Y_K = 1 | L_0 = (L_0^*, l_{\sigma(i)}), A_0 = a_{\sigma(i)})]\end{array}} \\
&= \frac{\begin{array}{c}\dfrac{\Pr(Y_K = 1 | L_0 = (L_0^*, l_0), A_0 = a_0)}{1 - \Pr(Y_K = 1 | L_0 = (L_0^*, l_0), A_0 = a_0)} \\ \times \displaystyle\prod_{i=0}^{M} [1 - \Pr(Y_K = 1 | L_0 = (L_0^*, l_i), A_0 = a_i)]\end{array}}{\begin{array}{c}\displaystyle\sum_{\sigma} \frac{\Pr(Y_K = 1 | L_0 = (L_0^*, l_{\sigma(0)}), A_0 = a_{\sigma(0)})}{1 - \Pr(Y_K = 1 | L_0 = (L_0^*, l_{\sigma(0)}), A_0 = a_{\sigma(0)})} \\ \times \displaystyle\prod_{i=0}^{M} [1 - \Pr(Y_K = 1 | L_0 = (L_0^*, l_{\sigma(i)}), A_0 = a_{\sigma(i)})]\end{array}} \\
&= \frac{\dfrac{\Pr(Y_K = 1 | L_0 = (L_0^*, l_0), A_0 = a_0)}{1 - \Pr(Y_K = 1 | L_0 = (L_0^*, l_0), A_0 = a_0)}}{\displaystyle\sum_{\sigma} \frac{\Pr(Y_K = 1 | L_0 = (L_0^*, l_{\sigma(0)}), A_0 = a_{\sigma(0)})}{1 - \Pr(Y_K = 1 | L_0 = (L_0^*, l_{\sigma(0)}), A_0 = a_{\sigma(0)})}} \\
&\propto \frac{\dfrac{\Pr(Y_K = 1 | L_0 = (L_0^*, l_0), A_0 = a_0)}{1 - \Pr(Y_K = 1 | L_0 = (L_0^*, l_0), A_0 = a_0)}}{\displaystyle\sum_{i=0}^{M} \frac{\Pr(Y_K = 1 | L_0 = (L_0^*, l_i), A_0 = a_i)}{1 - \Pr(Y_K = 1 | L_0 = (L_0^*, l_i), A_0 = a_i)}} \\
&= \frac{\dfrac{\operatorname{expit}\{\alpha + \sum_{j=1}^{p} f_j(a_0, L_0^*, l_0)\beta_j\}}{1 - \operatorname{expit}\{\alpha + \sum_{j=1}^{p} f_j(a_0, L_0^*, l_0)\beta_j\}}}{\displaystyle\sum_{i=0}^{M} \frac{\operatorname{expit}\{\alpha + \sum_{j=1}^{p} f_j(a_i, L_0^*, l_i)\beta_j\}}{1 - \operatorname{expit}\{\alpha + \sum_{j=1}^{p} f_j(a_i, L_0^*, l_i)\beta_j\}}} \\
&= \frac{\exp\left[\sum_{j=1}^{p} f_j(a_0, L_0^*, l_0)\beta_j\right]}{\sum_{i=0}^{M} \exp\left[\sum_{j=1}^{p} f_j(a_i, L_0^*, l_i)\beta_j\right]} \\
&= \left( \sum_{i=0}^{M} \exp\left[ \sum_{j=1}^{p} [f_j(a_i, L_0^*, l_i) - f_j(a_0, L_0^*, l_0)]\beta_j \right] \right)^{-1} \\
&= \left( 1 + \sum_{i=1}^{M} \exp\left[ \sum_{j=1}^{p} [f_j(a_i, L_0^*, l_i) - f_j(a_0, L_0^*, l_0)]\beta_j \right] \right)^{-1} .
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\sum_{i=1}^{m} X_{i,j} g(X_i, \tilde{\beta})}{1+\sum_{i=1}^{m} g(X_i, \tilde{\beta})}\middle| L_0^* = l^*, Y_K = 1, M = m\right] \\
&\quad \propto \sum_{l_0,\dots,l_m} \sum_{a_0,\dots,a_m} \frac{\begin{array}{r}\sum_{i=1}^{m}[f_k(a_i, l^*, l_i) - f_k(a_0, l^*, l_0)] \\ \times \exp\left\{ \sum_{k=1}^{p}[f_k(a_i, l^*, l_i) - f_k(a_0, l^*, l_0)]\tilde{\beta}_k\right\}\end{array}}{1+\sum_{i=1}^{m} \exp\left\{ \sum_{k=1}^{p}[f_k(a_i, l^*, l_i) - f_k(a_0, l^*, l_0)]\tilde{\beta}_k\right\}} \\
&\qquad \times \frac{1}{1+\sum_{i=1}^{m} \exp\left\{ \sum_{k=1}^{p}[f_k(a_i, l^*, l_i) - f_k(a_0, l^*, l_0)]\beta_k\right\}} \\
&\qquad \times h(a_0, ..., a_M, l_0, ..., l_M) \\
&\quad \propto \sum_{l_0,\dots,l_m} \sum_{a_0,\dots,a_m} h(a_0, ..., a_M, l_0, ..., l_M) \\
&\qquad \times \sum_{i=1}^{m}[f_k(a_i, l^*, l_i) - f_k(a_0, l^*, l_0)] \\
&\qquad \times \frac{\exp\left\{ \sum_{k=1}^{p}[f_k(a_i, l^*, l_i) - f_k(a_0, l^*, l_0)]\tilde{\beta}_k\right\}}{1+\sum_{i=1}^{m} \exp\left\{ \sum_{k=1}^{p}[f_k(a_i, l^*, l_i) - f_k(a_0, l^*, l_0)]\tilde{\beta}_k\right\}} \\
&\qquad \times \frac{1}{1+\sum_{i=1}^{m} \exp\left\{ \sum_{k=1}^{p}[f_k(a_i, l^*, l_i) - f_k(a_0, l^*, l_0)]\beta_k\right\}} \\
&\quad \propto \sum_{\{(l_0,a_0),\dots,(l_m,a_M)\}} h(a_0, ..., a_M, l_0, ..., l_M) \\
&\qquad \times \sum_{u=1}^{m}\sum_{i=1}^{m}[f_k(a_i, l^*, l_i) - f_k(a_u, l^*, l_u)] \\
&\qquad \times \frac{\exp\left\{ \sum_{k=1}^{p}[f_k(a_i, l^*, l_i) - f_k(a_u, l^*, l_u)]\tilde{\beta}_k\right\}}{1+\sum_{i=1}^{m} \exp\left\{ \sum_{k=1}^{p}[f_k(a_i, l^*, l_i) - f_k(a_u, l^*, l_u)]\tilde{\beta}_k\right\}} \\
&\qquad \times \frac{1}{1+\sum_{i=1}^{m} \exp\left\{ \sum_{k=1}^{p}[f_k(a_i, l^*, l_i) - f_k(a_u, l^*, l_u)]\beta_k\right\}} \\
&\quad = \sum_{\{(l_0,a_0),\dots,(l_m,a_M)\}} h(a_0, ..., a_M, l_0, ..., l_M) \\
&\qquad \times \sum_{u=1}^{m}\sum_{i=1}^{m}[f_k(a_i, l^*, l_i) - f_k(a_u, l^*, l_u)] \frac{\exp\left\{ \sum_{k=1}^{p} f_k(a_i, l^*, l_i)\tilde{\beta}_k\right\}}{\sum_{i=0}^{m} \exp\left\{ \sum_{k=1}^{p} f_k(a_i, l^*, l_i)\tilde{\beta}_k\right\}} \\
&\qquad \times \frac{\exp\left\{ \sum_{k=1}^{p} f_k(a_u, l^*, l_u)\beta_k\right\}}{\sum_{i=0}^{m} \exp\left\{ \sum_{k=1}^{p} f_k(a_i, l^*, l_i)\beta_k\right\}}
\end{aligned}
$$

$$\begin{aligned}
&= \sum_{\{(l_0,a_0),...,(l_m,a_M)\}} h(a_0,...,a_M,l_0,...,l_M) \\
&\quad \times \Bigg[ \sum_{u,i\in\{1,...,m\}:i>u} [f_k(a_i,l^*,l_i) - f_k(a_u,l^*,l_u)] \\
&\quad \times \frac{\exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\tilde{\beta}_k\}}{\sum_{i=0}^m \exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\tilde{\beta}_k\}} \frac{\exp\{\sum_{k=1}^p f_k(a_u,l^*,l_u)\beta_k\}}{\sum_{i=0}^m \exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\beta_k\}} \\
&\quad + \sum_{u,i\in\{1,...,m\}:i<u} [f_k(a_i,l^*,l_i) - f_k(a_u,l^*,l_u)] \\
&\quad \times \frac{\exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\tilde{\beta}_k\}}{\sum_{i=0}^m \exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\tilde{\beta}_k\}} \frac{\exp\{\sum_{k=1}^p f_k(a_u,l^*,l_u)\beta_k\}}{\sum_{i=0}^m \exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\beta_k\}} \Bigg] \\
&= \sum_{\{(l_0,a_0),...,(l_m,a_M)\}} h(a_0,...,a_M,l_0,...,l_M) \\
&\quad \times \Bigg[ \sum_{u,i\in\{1,...,m\}:i>u} [f_k(a_i,l^*,l_i) - f_k(a_u,l^*,l_u)] \\
&\quad \times \frac{\exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\tilde{\beta}_k\}}{\sum_{i=0}^m \exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\tilde{\beta}_k\}} \frac{\exp\{\sum_{k=1}^p f_k(a_u,l^*,l_u)\beta_k\}}{\sum_{i=0}^m \exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\beta_k\}} \\
&\quad - \sum_{u,i\in\{1,...,m\}:i>u} [f_k(a_i,l^*,l_i) - f_k(a_u,l^*,l_u)] \\
&\quad \times \frac{\exp\{\sum_{k=1}^p f_k(a_u,l^*,l_u)\tilde{\beta}_k\}}{\sum_{i=0}^m \exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\tilde{\beta}_k\}} \frac{\exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\beta_k\}}{\sum_{i=0}^m \exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\beta_k\}} \Bigg] \\
&= \sum_{\{(l_0,a_0),...,(l_m,a_M)\}} h(a_0,...,a_M,l_0,...,l_M) \\
&\quad \times \sum_{u,i\in\{1,...,m\}:i>u} [f_k(a_i,l^*,l_i) - f_k(a_u,l^*,l_u)] \\
&\quad \times \Bigg[ \frac{\exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\tilde{\beta}_k\}}{\sum_{i=0}^m \exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\tilde{\beta}_k\}} \frac{\exp\{\sum_{k=1}^p f_k(a_u,l^*,l_u)\beta_k\}}{\sum_{i=0}^m \exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\beta_k\}} \\
&\quad - \frac{\exp\{\sum_{k=1}^p f_k(a_u,l^*,l_u)\tilde{\beta}_k\}}{\sum_{i=0}^m \exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\tilde{\beta}_k\}} \frac{\exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\beta_k\}}{\sum_{i=0}^m \exp\{\sum_{k=1}^p f_k(a_i,l^*,l_i)\beta_k\}} \Bigg],
\end{aligned}$$

which is clearly zero when $\tilde{\beta} = \beta$. If follows that

$$\frac{\partial}{\partial \tilde{\beta}_k} f(\tilde{\beta}) = \mathbb{E}\left[ \frac{\sum_{i=1}^M X_{i,k} g(X_i,\tilde{\beta})}{1 + \sum_{i=1}^m g(X_i,\tilde{\beta})} \middle| Y_K = 1, M > 0 \right] = 0$$

for all $k = 1, ..., p$ if and only if $\tilde{\beta} = \beta$. $\square$

# 11

# On selecting optimal subgroups for treatment using many covariates

Bas B. L. Penning de Vries
Rolf H. H. Groenwold
Alex Luedtke

*Epidemiology* 2020; 31(4): e33–e34

Optimal sugroups

In a recent publication, VanderWeele et al. (2019) considered the task of finding a treatment subgroup that maximizes the mean potential outcome. They showed that the task can sometimes be considerably simplified by deriving optimal treatment assignment rules of a simple form: assign treatment in a greedy fashion to all individuals with the next largest benefit (i.e., the difference in potential outcome means given covariates) or the next highest benefit–cost ratio (with cost being a positive function of baseline covariates) until the resource or cost constraint, respectively, is exceeded. As they state in their supplementary material, the optimality of the rules relies critically on the assumption that there are no ties between individuals. Although tied treatment effects or benefit–cost ratios may occur with many covariates, they are perhaps more realistic when few and only discrete baseline variables are considered to define treatment rules.

Consider for example the setting of Table 11.1 and suppose that the total cost may not exceed 130. According to the rule of VanderWeele et al. (2019), individuals in the first stratum should be assigned treatment. Because the

**Table 11.1:** Characteristics of hypothetical population of size 100 with baseline covariates forming five strata.

| | Stratum | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Number of individuals | 25 | 20 | 10 | 15 | 30 |
| Conditional mean potential outcome | | | | | |
| – under no treatment | −5 | 4 | 0 | −5 | −5 |
| – under treatment | 15 | 20 | 20 | 5 | −15 |
| Cost of treatment per individual | 4 | 4 | 5 | 10 | 10 |
| Benefit–cost ratio | 5 | 4 | 4 | 1 | −1 |

If those and only those in stratum 1 are treated, the total cost is $25 \times 4 = 100$ and the mean potential outcome is $(25\times15+20\times4+10\times0+15\times-5+30\times-5)/(25+20+10+15+30) = 230/100 = 2.3$. If those and only those patients in strata 2 and 3 are treated, the total cost is $20 \times 4 + 10 \times 5 = 130$ and the mean potential outcome is $(25\times-5+20\times20+10\times20+15\times-5+30\times-5)/(25+20+10+15+30) = 250/100 = 2.5$. If patients in stratum 1 are treated with probability 1, patients in strata 2 and 3 with probability 3/13, and the rest with probability 0, the expected total cost is $25\times4+(3/13)\times20\times4+(3/13)\times10\times5 = 130$ and the mean potential outcome is $(25 \times 15 + (3/13) \times 20 \times 20 + (10/13) \times 20 \times 4 + (3/13) \times 10 \times 20 + (10/13) \times 10 \times 0 + 15 \times -5 + 30 \times -5)/(25 + 20 + 10 + 15 + 30) = 350/100 = 3.5$.

presented rules assign treatment to either all or no individuals in any given stratum, no more individuals can be selected without violating the cost constraint. This rule yields a mean potential outcome of 2.3. However, because of ties, a better rule that likewise selects either all or no individuals of a stratum, does exist: assign treatment to strata 2 and 3 (with a mean potential outcome of 2.5). Thus, in the presence of ties, the optimal rule need not be greedy (see also the literature on the classic knapsack problem; e.g., Korte and Vygen, 2008).

We note that a better rule may be obtained by augmenting our data with a sequence of independent, possibly unfair, coin tosses. As shown in the eAppendix (but see also Luedtke and van der Laan, 2016), maximizing the mean potential outcome across rules of this kind is achieved in the cost-constrained setting by treating those with a benefit–cost ratio strictly greater than some positive constant and a random selection of those with a benefit–cost ratio that equals that constant. For our example, this means treating all members of stratum 1 as well as those members of strata 2 and 3 whose independent coin toss, with probability 3/13 of showing heads, results in heads (mean potential outcome: 3.5).

It seems unlikely that these treatment rules would be implemented via biased coin tosses in real-world settings. If resources are made available in a single batch, one could calculate the amount of resources that would need to be allocated to the "always-treat" portion of the population, reserve this portion of resources for always-treat individuals, and then allocate the remainder to the "sometimes-treat" portion of the population on a first-come, first-serve basis until that portion of resources runs out. Bias could however be introduced by doing this, for example, when sometimes-treat individuals who visit the clinic more frequently are systematically less (or more) likely to benefit from treatment. However, there may be ways to account for this (e.g., by including frequency of visits as a covariate).

Finally, we add that with multiple treatment levels and cost constraints, mean potential outcomes need not be optimized by the greedy approach of assigning to subjects the treatment level with the highest benefit–cost ratio above or at treatment level-specific thresholds (to satisfy cost constraints), even if the observed data are augmented with a sequence of independent coin tosses (Supplementary Material). Regardless of the form the rule should take, however, we encourage researchers to follow VanderWeele et al. (2019) in taking a more formal approach to "precision medicine" with clearly specified objectives, so that the optimal rule form may be derived and estimation strategies be evaluated.

## References

VanderWeele, T. J., A. R. Luedtke, M. J. van der Laan, and R. C. Kessler (2019): “Selecting optimal subgroups for treatment using many covariates,” *Epidemiology*, 30, 334–341.

Korte, B., and J. Vygen (2008): *Combinatorial Optimization: Theory and Algorithms*, Springer, fourth edition.

Luedtke, A. R., and M. J. van der Laan (2016): “Optimal dynamic treatments in resource-limited settings,” *International Journal of Biostatistics*, 12, 283–303.

## Supplementary Material

In what follows, we denote by $I$ the indicator function and by $Y^a$ the counterfactual or potential outcome that would be realised if, possibly contrary to fact, $A$ were set to $a$. Superscripts are reserved for assigned treatment levels rather than powers. For example, $Y^{I(S)}$ is the counterfactual outcome $Y^1$ if statement $S$ is true and is $Y^0$ otherwise. We consider treatment assignment rules that map the vector $X$ of covariate vector $L$ and an error term $\varepsilon$ to the value of 0 or 1. We generally require that $\varepsilon$ be independent of $(Y^1 - Y^0, L)$ and uniformly distributed between 0 and 1, so that for fixed $p \in [0, 1]$, $I(\varepsilon < p)$ takes the Bernoulli distribution with parameter $p$ and, as such, behaves like an independent (unfair) coin toss.

**Lemma 11.1.** *Let $\mathcal{X}$ be the support of $X := (L, \varepsilon)$ and suppose that $(Y^1 - Y^0) \perp\!\!\!\perp \varepsilon | L$. If $\mathcal{X}_0 \subseteq \mathcal{X}_1 \subseteq \mathcal{X}$ such that $(L, \varepsilon) \in \mathcal{X}_1 \Rightarrow \mathbb{E}[Y^1 - Y^0 | L] > 0$, then $\mathbb{E}\big[Y^{I(X \in \mathcal{X}_1)}\big] \geq \mathbb{E}\big[Y^{I(X \in \mathcal{X}_0)}\big]$. Also, for all $\mathcal{X}' \subseteq \mathcal{X}$, we have $\mathbb{E}\big[Y^{I(X \in \mathcal{X}' \wedge \mathbb{E}[Y^1 - Y^0 | L] > 0)}\big] \geq \mathbb{E}\big[Y^{I(X \in \mathcal{X}')}\big]$.*

*Proof.* Define $\mathcal{X}_0$ and $\mathcal{X}_1$ as indicated above, so that

$$\begin{aligned}
&\mathbb{E}\big[Y^{I(X \in \mathcal{X}_1)}\big] \\
&\quad = \mathbb{E}\big[Y^{I(X \in \mathcal{X}_0 \vee X \in \mathcal{X}_1 \backslash \mathcal{X}_0)}\big] \\
&\quad = \mathbb{E}[Y^0] + \mathbb{E}[(Y^1 - Y^0) I(X \in \mathcal{X}_0 \vee X \in \mathcal{X}_1 \backslash \mathcal{X}_0)] \\
&\quad = \mathbb{E}[Y^0] + \mathbb{E}[(Y^1 - Y^0) I(X \in \mathcal{X}_0)] + \mathbb{E}[(Y^1 - Y^0) I(X \in \mathcal{X}_1 \backslash \mathcal{X}_0)] \\
&\quad = \mathbb{E}\big[Y^{I(X \in \mathcal{X}_0)}\big] + \mathbb{E}[(Y^1 - Y^0) I(X \in \mathcal{X}_1 \backslash \mathcal{X}_0)].
\end{aligned}$$

If $\Pr(X \in \mathcal{X}_1 \backslash \mathcal{X}_0) > 0$, then

$$\begin{aligned}
&\mathbb{E}[(Y^1 - Y^0) I(X \in \mathcal{X}_1 \backslash \mathcal{X}_0)] \\
&\quad = \mathbb{E}[Y^1 - Y^0 | X \in \mathcal{X}_1 \backslash \mathcal{X}_0] \Pr(X \in \mathcal{X}_1 \backslash \mathcal{X}_0) \\
&\quad = \mathbb{E}\{\mathbb{E}[Y^1 - Y^0 | L, \varepsilon] | X \in \mathcal{X}_1 \backslash \mathcal{X}_0\} \Pr(X \in \mathcal{X}_1 \backslash \mathcal{X}_0) \\
&\quad = \mathbb{E}\{\mathbb{E}[Y^1 - Y^0 | L] | X \in \mathcal{X}_1 \backslash \mathcal{X}_0\} \Pr(X \in \mathcal{X}_1 \backslash \mathcal{X}_0),
\end{aligned}$$

which is strictly positive, because the inner expectation is strictly positive on (any subset of) $\mathcal{X}_1$. Also, if $\Pr(X \in \mathcal{X}_1 \backslash \mathcal{X}_0) = 0$, then $\mathbb{E}[(Y^1 - Y^0) I(X \in \mathcal{X}_1 \backslash \mathcal{X}_0)] = 0$. In either case, $\mathbb{E}\big[Y^{I(X \in \mathcal{X}_1)}\big] \geq \mathbb{E}\big[Y^{I(X \in \mathcal{X}_0)}\big]$.

As for the last statement, fix some $\mathcal{X}' \subseteq \mathcal{X}$, let $\mathcal{X}'' = \{X \subseteq \mathcal{X} : \mathbb{E}[Y^1 - Y^0 | L] > 0\}$ and observe

$$\mathbb{E}\big[Y^{I(X \in \mathcal{X}' \wedge \mathbb{E}[Y^1 - Y^0 | L] > 0)}\big]$$

$$
\begin{aligned}
&= \mathbb{E}\big[Y^{1-I(X\in\mathcal{X}\backslash\mathcal{X}' \vee X\in\mathcal{X}\backslash\mathcal{X}'')}\big] \\
&= \mathbb{E}[Y^1] + \mathbb{E}[(Y^0 - Y^1)I(X \in \mathcal{X}\backslash\mathcal{X}' \vee X \in \mathcal{X}\backslash\mathcal{X}'')] \\
&= \mathbb{E}[Y^1] + \mathbb{E}[(Y^0 - Y^1)I(X \in \mathcal{X}\backslash\mathcal{X}' \vee X \in (\mathcal{X}\backslash\mathcal{X}'')\backslash(\mathcal{X}\backslash\mathcal{X}'))] \\
&= \mathbb{E}[Y^1] + \mathbb{E}[(Y^0 - Y^1)I(X \in \mathcal{X}\backslash\mathcal{X}')] + \mathbb{E}[(Y^0 - Y^1)I(X \in \mathcal{X}'\backslash\mathcal{X}'')] \\
&= \mathbb{E}\big[Y^{X\in\mathcal{X}'}\big] + \mathbb{E}[(Y^0 - Y^1)I(X \in \mathcal{X}'\backslash\mathcal{X}'')]
\end{aligned}
$$

with $\mathbb{E}[(Y^0 - Y^1)I(X \in \mathcal{X}'\backslash\mathcal{X}'')] = 0$ if $\Pr(X \in \mathcal{X}'\backslash\mathcal{X}'') = 0$ and, if $\Pr(X \in \mathcal{X}'\backslash\mathcal{X}'') > 0$,

$$
\begin{aligned}
&\mathbb{E}[(Y^0 - Y^1)I(X \in \mathcal{X}'\backslash\mathcal{X}'')] \\
&\qquad = \mathbb{E}[Y^0 - Y^1 | X \in \mathcal{X}'\backslash\mathcal{X}''] \Pr(X \in \mathcal{X}'\backslash\mathcal{X}'') \\
&\qquad = \mathbb{E}\{-\mathbb{E}[Y^1 - Y^0 | L, \varepsilon] | X \in \mathcal{X}'\backslash\mathcal{X}''\} \Pr(X \in \mathcal{X}'\backslash\mathcal{X}'') \\
&\qquad = \mathbb{E}\{-\mathbb{E}[Y^1 - Y^0 | L] | X \in \mathcal{X}'\backslash\mathcal{X}''\} \Pr(X \in \mathcal{X}'\backslash\mathcal{X}'').
\end{aligned}
$$

Because the inner expectation is strictly negative on (any subset of) $\mathcal{X}\backslash\mathcal{X}''$, we have $\mathbb{E}[(Y^0 - Y^1)I(X \in \mathcal{X}'\backslash\mathcal{X}'')] > 0$ if $\Pr(X \in \mathcal{X}'\backslash\mathcal{X}'') > 0$. Hence, $\mathbb{E}\big[Y^{I(X\in\mathcal{X}' \wedge \mathbb{E}[Y^1-Y^0|L]>0)}\big] \geq \mathbb{E}\big[Y^{X\in\mathcal{X}'}\big]$, as desired. □

**Lemma 11.2.** *Let $\mathcal{X}$ be the support of $X := (L, \varepsilon)$ and let $Cost$ be a deterministic, positive function of $L$ such that $\mathbb{E}[Cost(L)] \in \mathbb{R}$. For some positive real $\tau \leq \mathbb{E}[Cost(L)]$, define $\mathcal{G}$ to be the set of all deterministic functions $g : \mathcal{X} \to \{0,1\}$ such that $\mathbb{E}[Cost(L)g(X)] = \tau$. Suppose that $\varepsilon \perp\!\!\!\perp (Y^1 - Y^0, L)$, that $\varepsilon \sim \text{Uniform}[0,1]$ and that $\mathbb{E}[Y^1 - Y^0|L]$ is defined almost surely. Let $h(L) = \mathbb{E}\big[Y^a - Y^0|L\big]/Cost(L)$ and define $g^*$ such that*

$$
g^*((L,\varepsilon)) = \begin{cases} 1 & \text{if } h(L) > k, \\ 1 & \text{if } h(L) = k \wedge \varepsilon < p, \\ 0 & \text{if } h(L) < k \end{cases}
$$

*for all $(L,\varepsilon) \in \mathcal{X}$, and let $k = -\infty$ denote that $h(L) > k$ is necessarily true. Then, there exist $k \in \mathbb{R} \cup \{-\infty\}$ and $p \in [0,1]$ such that $g^* \in \mathcal{G}$.*

*Proof.* If $\tau = \mathbb{E}[Cost(L)]$, then letting $k = -\infty$ and $p = 0$ gives the result. So assume that $\tau < \mathbb{E}[Cost(L)]$.

Now, let

$$
f : k \mapsto \mathbb{E}\big[Cost(L)I(h(L) \geq k)\big]
$$

and $K = \{k \in \mathbb{R} : f(k) < \tau\}$.

Note that $f$ is upper semi-continuous (which can be seen to hold because $f$ is left continuous with right limits and monotonically non-increasing). Since upper semi-continuity of $f$ implies $\{x \in \mathbb{R} : f(x) < y\}$ is open for every $y \in \mathbb{R}$, we see that $\mathbb{R}\backslash K$ is closed.

To see that $\mathbb{R}\backslash K$ is nonempty, note that, by the dominated convergence theorem, $\lim_{k\to-\infty} f(k) = \mathbb{E}[Cost(L)] > \tau$. Hence, there exists $k_0 > -\infty$ such that $f(k_0) \geq \tau$, which in turn implies that $\mathbb{R}\backslash K$ is non-empty. Moreover, $\lim_{k\to\infty} f(k) = 0 < \tau$, and so there exists a $k_1$ such that $f(k_1) < \tau$. Hence, $\mathbb{R}\backslash K$ is bounded above.

Since $\mathbb{R}\backslash K$ is closed, non-empty, and bounded above, we see that $k := \sup \mathbb{R}\backslash K$ belongs to $\mathbb{R}\backslash K$, which implies that $f(k) \geq \tau$. The proof is complete if we can show that there exists a $p \in [0,1]$ such that $\tau = \mathbb{E}[Cost(L)g^*((L,\varepsilon))]$, where we note that $g^*$ depends on the choice of $p$. To see that this is the case, first note that

$$\begin{aligned}\mathbb{E}[Cost(L)g^*((L,\varepsilon))] &= \mathbb{E}\big[Cost(L)I(h(L) > k)\big] + p\mathbb{E}\big[Cost(L)I(h(L) = k)\big] \\ &= (1-p)\mathbb{E}\big[Cost(L)I(h(L) > k)\big] + pf(k) \\ &= (1-p)\lim_{k'\downarrow k} f(k') + pf(k).\end{aligned}$$

Now, for any $k' \geq k$, it holds that $k' \in K$, implying that $f(k') < \tau$. Hence, $\lim_{k'\downarrow k} f(k') \leq \tau$. Combining this fact with the fact that $f(k) \geq \tau$, we see that there exists a $p \in [0,1]$ such that $(1-p)\lim_{k'\downarrow k} f(k') + pf(k) = \tau$. This completes the proof. □

**Remark.** The constraint $\tau \leq \mathbb{E}[Cost(L)]$ in Lemma 11.2 is weaker than, and so may me replaced with, $\tau \leq \mathbb{E}\big[Cost(L)I\big(\mathbb{E}[Y^1 - Y^0|L] > 0\big)\big]$.

**Theorem 11.1.** *Consider some positive real $\tau$. In the setting of Lemma 11.2, except with $\mathcal{G}$ defined to be the set of all deterministic functions $g : \mathcal{X} \to \{0,1\}$ such that $\mathbb{E}[Cost(L)g(X)] \leq \tau$, (i) there exist $k \in (0,\infty)$ and $p \in [0,1]$ such that $g^* \in \mathcal{G}$ and (ii)*

$$g^* \in \operatorname*{arg\,max}_{g\in\mathcal{G}} \mathbb{E}\big[Y^{g(X)}\big].$$

*Proof.* Since $Y^{g(X)} = Y^0 + (Y^1 - Y^0)g(X)$ by consistency, we have

$$\begin{aligned}\mathbb{E}\big[Y^{g(X)}\big] &= \mathbb{E}\big[Y^0 + (Y^1 - Y^0)g(X)\big] \\ &= \mathbb{E}\big[Y^0\big] + \mathbb{E}\big[(Y^1 - Y^0)g(X)\big] \\ &= \mathbb{E}\big[Y^0\big] + \mathbb{E}\{\mathbb{E}[(Y^1 - Y^0)g(X)|g(X)]\}\end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}\big[Y^0\big] + \mathbb{E}\big[Y^1 - Y^0 | g(X) = 1\big]\mathbb{E}[g(X)] \\
&= \mathbb{E}\big[Y^0\big] + \frac{\mathbb{E}\big[Y^1 - Y^0 | g(X) = 1\big]}{\mathbb{E}[Cost(L)|g(X) = 1]}\mathbb{E}[Cost(L)g(X)].
\end{aligned}$$

Lemma 11.1 suggests choosing among all $g \in \mathcal{G}$ such that $\mathbb{E}[Cost(L)g(X)] = \min\big\{\tau, \mathbb{E}\big[Cost(L)I\big(\mathbb{E}[Y^1 - Y^0|L] > 0)\big]\big\}$. Let $\mathcal{G}'$ be the set of all such $g$. Since $\mathbb{E}\big[Y^0\big]$ and $\mathbb{E}[Cost(L)g(X)]$ are invariant under changes in $g \in \mathcal{G}'$,

$$\arg\max_{g \in \mathcal{G}} \mathbb{E}\big[Y^{g(X)}\big] \supseteq \arg\max_{g \in \mathcal{G}'} \frac{\mathbb{E}\big[Y^1 - Y^0 | g(X) = 1\big]}{\mathbb{E}[Cost(L)|g(X) = 1]}.$$

Part (i) now follows from Lemma 11.2. In the remainder of this proof, we show that (ii) holds also. It suffices to show that

$$g^* \in \arg\max_{g \in \mathcal{G}'} \frac{\mathbb{E}\big[Y^1 - Y^0 | g(X) = 1\big]}{\mathbb{E}[Cost(L)|g(X) = 1]}.$$

To show that the above expression is true, consider first any non-empty $\mathcal{X}_0, \mathcal{X}_1 \subseteq \mathcal{X}$ such that $\mathbb{E}[Cost(L)I(X \in \mathcal{X}_0)] = \mathbb{E}[Cost(L)I(X \in \mathcal{X}_1)] = \tau'$ for some $\tau' \in \mathbb{R}_+$. It holds that

$$\begin{aligned}
\tau' &= \mathbb{E}[Cost(L)I(X \in \mathcal{X}_0)] \\
&= \mathbb{E}[Cost(L)I(X \in \mathcal{X}_0 \cap \mathcal{X}_1) + Cost(L)I(X \in \mathcal{X}_0 \backslash \mathcal{X}_1)] \\
&= \mathbb{E}[Cost(L)|X \in \mathcal{X}_0 \cap \mathcal{X}_1] \Pr(X \in \mathcal{X}_0 \cap \mathcal{X}_1) \\
&\qquad + \mathbb{E}[Cost(L)|X \in \mathcal{X}_0 \backslash \mathcal{X}_1] \Pr(X \in \mathcal{X}_0 \backslash \mathcal{X}_1)
\end{aligned}$$

and, similarly,

$$\begin{aligned}
\tau' &= \mathbb{E}[Cost(L)|X \in \mathcal{X}_0 \cap \mathcal{X}_1] \Pr(X \in \mathcal{X}_0 \cap \mathcal{X}_1) \\
&\qquad + \mathbb{E}[Cost(L)|X \in \mathcal{X}_1 \backslash \mathcal{X}_0] \Pr(X \in \mathcal{X}_1 \backslash \mathcal{X}_0),
\end{aligned}$$

so that $\mathbb{E}[Cost(L)|X \in \mathcal{X}_0 \backslash \mathcal{X}_1] \Pr(X \in \mathcal{X}_0 \backslash \mathcal{X}_1) = \mathbb{E}[Cost(L)|X \in \mathcal{X}_1 \backslash \mathcal{X}_0] \Pr(X \in \mathcal{X}_1 \backslash \mathcal{X}_0)$. Therefore, there exist $a \in \mathbb{R}$ and $b, c \in \mathbb{R}_+ \cup \{0\}$ such that $b + c \neq 0$ and for all $i \in \{0, 1\}$,

$$\begin{aligned}
&\frac{\mathbb{E}\big[Y^1 - Y^0 | X \in \mathcal{X}_i\big]}{\mathbb{E}[Cost(L)|X \in \mathcal{X}_i]} \\
&\quad = \frac{\begin{array}{l}\mathbb{E}\big[Y^1 - Y^0 | X \in \mathcal{X}_i \cap \mathcal{X}_{1-i}\big] \Pr(X \in \mathcal{X}_{1-i} | X \in \mathcal{X}_i) \\ \quad + \mathbb{E}\big[Y^1 - Y^0 | X \in \mathcal{X}_i \backslash \mathcal{X}_{1-i}\big] \Pr(X \notin \mathcal{X}_{1-i} | X \in \mathcal{X}_i)\end{array}}{\begin{array}{l}\mathbb{E}[Cost(L)|X \in \mathcal{X}_i \cap \mathcal{X}_{1-i}] \Pr(X \in \mathcal{X}_{1-i} | X \in \mathcal{X}_i) \\ \quad + \mathbb{E}[Cost(L)|X \in \mathcal{X}_i \backslash \mathcal{X}_{1-i}] \Pr(X \notin \mathcal{X}_{1-i} | X \in \mathcal{X}_i)\end{array}}
\end{aligned}$$

$$= \frac{\begin{array}{l}\mathbb{E}[Y^1 - Y^0|X \in \mathcal{X}_i \cap \mathcal{X}_{1-i}]\,\Pr(X \in \mathcal{X}_{1-i} \cap \mathcal{X}_i) \\ \quad + \mathbb{E}[Y^1 - Y^0|X \in \mathcal{X}_i \backslash \mathcal{X}_{1-i}]\,\Pr(X \in \mathcal{X}_i \backslash \mathcal{X}_{1-i})\end{array}}{\begin{array}{l}\mathbb{E}[Cost(L)|X \in \mathcal{X}_i \cap \mathcal{X}_{1-i}]\,\Pr(X \in \mathcal{X}_{1-i} \cap \mathcal{X}_i) \\ \quad + \mathbb{E}[Cost(L)|X \in \mathcal{X}_i \backslash \mathcal{X}_{1-i}]\,\Pr(X \in \mathcal{X}_i \backslash \mathcal{X}_{1-i})\end{array}}$$

$$= \frac{a + \mathbb{E}[Y^1 - Y^0|X \in \mathcal{X}_i \backslash \mathcal{X}_{1-i}]\mathbb{E}[Cost(L)|X \in \mathcal{X}_i \backslash \mathcal{X}_{1-i}]^{-1}b}{c + b}.$$

This readily shows that

$$\begin{aligned} &\frac{\mathbb{E}[Y^1 - Y^0|X \in \mathcal{X}_0]}{\mathbb{E}[Cost(L)|X \in \mathcal{X}_0]} > \frac{\mathbb{E}[Y^1 - Y^0|X \in \mathcal{X}_1]}{\mathbb{E}[Cost(L)|X \in \mathcal{X}_1]} \\ &\iff \frac{\mathbb{E}[Y^1 - Y^0|X \in \mathcal{X}_0 \backslash \mathcal{X}_1]}{\mathbb{E}[Cost(L)|X \in \mathcal{X}_0 \backslash \mathcal{X}_1]} > \frac{\mathbb{E}[Y^1 - Y^0|X \in \mathcal{X}_1 \backslash \mathcal{X}_0]}{\mathbb{E}[Cost(L)|X \in \mathcal{X}_1 \backslash \mathcal{X}_0]} \end{aligned} \tag{11.1}$$

for any non-empty $\mathcal{X}_0, \mathcal{X}_1 \subseteq \mathcal{X}$ such that $\mathbb{E}[Cost(L)I(X \in \mathcal{X}_0)] = \mathbb{E}[Cost(L)I(X \in \mathcal{X}_1)] = \tau'$ for some $\tau' \in \mathbb{R}_+$.

Let $\mathcal{X}_0 = \{X \in \mathcal{X} : g^*(X) = 1\}$. Suppose, by way of contradiction, that there exists $\mathcal{X}_1$ such that $\mathbb{E}[Cost(L)I(X \in \mathcal{X}_0)] = \mathbb{E}[Cost(L)I(X \in \mathcal{X}_1)]$ and

$$\frac{\mathbb{E}[Y^1 - Y^0|X \in \mathcal{X}_0]}{\mathbb{E}[Cost(L)|X \in \mathcal{X}_0]} < \frac{\mathbb{E}[Y^1 - Y^0|X \in \mathcal{X}_1]}{\mathbb{E}[Cost(L)|X \in \mathcal{X}_1]},$$

so that, by (11.1),

$$\frac{\mathbb{E}[Y^1 - Y^0|X \in \mathcal{X}_0 \backslash \mathcal{X}_1]}{\mathbb{E}[Cost(L)|X \in \mathcal{X}_0 \backslash \mathcal{X}_1]} < \frac{\mathbb{E}[Y^1 - Y^0|X \in \mathcal{X}_1 \backslash \mathcal{X}_0]}{\mathbb{E}[Cost(L)|X \in \mathcal{X}_1 \backslash \mathcal{X}_0]}. \tag{11.2}$$

Sets $\mathcal{X}_0 \backslash \mathcal{X}_1$ and $\mathcal{X}_1 \backslash \mathcal{X}_0$ are disjoint and $\mathbb{E}[Cost(L)I(X \in \mathcal{X}_0 \backslash \mathcal{X}_1)] = \mathbb{E}[Cost(L)I(X \in \mathcal{X}_1 \backslash \mathcal{X}_0)]$. In addition, for all non-empty subsets $\mathcal{X}_0' \subseteq \mathcal{X}_0 \backslash \mathcal{X}_1$ and $\mathcal{X}_1' \subseteq \mathcal{X}_1 \backslash \mathcal{X}_0$, we have, by construction of $\mathcal{X}_0$ and disjointedness, that

$$\inf\left\{\frac{\mathbb{E}[Y^1 - Y^0|L]}{Cost(L)} : X \in \mathcal{X}_0'\right\} \geq \sup\left\{\frac{\mathbb{E}[Y^1 - Y^0|L]}{Cost(L)} : X \in \mathcal{X}_1'\right\}. \tag{11.3}$$

Let $f(L) = \mathbb{E}[Y^1 - Y^0|L]$ and $g(L) = Cost(L)$, so that $h(L) = f(L)/g(L)$, and observe that

$$\begin{aligned} \frac{\mathbb{E}[f(L)|X \in \mathcal{X}_0']}{\mathbb{E}[g(L)|X \in \mathcal{X}_0']} &= \mathbb{E}\left[\frac{f(L)}{g(L)} \frac{g(L)}{\mathbb{E}[g(L)|X \in \mathcal{X}_0']}\middle| X \in \mathcal{X}_0'\right] \\ &\geq \mathbb{E}\left[\inf\left\{\frac{f(L)}{g(L)} : X \in \mathcal{X}_0'\right\}\frac{g(L)}{\mathbb{E}[g(L)|X \in \mathcal{X}_0']}\middle| X \in \mathcal{X}_0'\right] \end{aligned}$$

$$= \inf\left\{\frac{f(L)}{g(L)} : X \in \mathcal{X}_0'\right\}\mathbb{E}\left[\frac{g(L)}{\mathbb{E}[g(L)|X \in \mathcal{X}_0']}\middle|X \in \mathcal{X}_0'\right]$$
$$= \inf\left\{h(L) : X \in \mathcal{X}_0'\right\}. \tag{11.4}$$

Similarly, we have

$$\frac{\mathbb{E}[f(L)|X \in \mathcal{X}_1']}{\mathbb{E}[g(L)|X \in \mathcal{X}_1']} \leq \sup\left\{h(L) : X \in \mathcal{X}_1'\right\}. \tag{11.5}$$

Taken together, (11.3), (11.4) and (11.5) imply

$$\frac{\mathbb{E}\{\mathbb{E}[Y^1 - Y^0|L]|X \in \mathcal{X}_0'\}}{\mathbb{E}[Cost(L)|X \in \mathcal{X}_0']} \geq \frac{\mathbb{E}\{\mathbb{E}[Y^1 - Y^0|L]|X \in \mathcal{X}_1'\}}{\mathbb{E}[Cost(L)|X \in \mathcal{X}_1']},$$

which, by assumption that $(Y^1 - Y^0, L) \perp\!\!\!\perp \varepsilon$ (and, in turn, $(Y^1 - Y^0) \perp\!\!\!\perp \varepsilon|L$ by weak union), implies

$$\frac{\mathbb{E}[Y^1 - Y^0|X \in \mathcal{X}_0']}{\mathbb{E}[Cost(L)|X \in \mathcal{X}_0']} \geq \frac{\mathbb{E}[Y^1 - Y^0|X \in \mathcal{X}_1']}{\mathbb{E}[Cost(L)|X \in \mathcal{X}_1']}.$$

In particular, this implies

$$\frac{\mathbb{E}[Y^1 - Y^0|X \in \mathcal{X}_0\backslash\mathcal{X}_1]}{\mathbb{E}[Cost(L)|X \in \mathcal{X}_0\backslash\mathcal{X}_1]} \geq \frac{\mathbb{E}[Y^1 - Y^0|X \in \mathcal{X}_1\backslash\mathcal{X}_0]}{\mathbb{E}[Cost(L)|X \in \mathcal{X}_1\backslash\mathcal{X}_0]}.$$

However, in view of (11.2), this poses a contradiction. Hence, for all $g \in \mathcal{G}'$, we have

$$\frac{\mathbb{E}[Y^1 - Y^0|g^*(X) = 1]}{\mathbb{E}[Cost(L)|g^*(X) = 1]} \geq \frac{\mathbb{E}[Y^1 - Y^0|g(X) = 1]}{\mathbb{E}[Cost(L)|g(X) = 1]},$$

so that $g^* \in \arg\max\limits_{g \in \mathcal{G}} \mathbb{E}\big[Y^{g(X)}\big]$, as desired. □

The counterexample to the following proposition suggests that the a greedy approach need not optimize mean potential outcomes with multiple treatment levels and cost or resource constraints.

**Proposition.** *Let $\mathcal{A}$ be a finite set that includes $0$ and denote by $\mathcal{X}$ the support of $X := (L, \varepsilon)$. For $a \in \mathcal{A}\backslash\{0\}$, let $Cost_a$ be a deterministic, positive function of $L$ such that $\mathbb{E}[Cost_a(L)] \in \mathbb{R}$. Let $I$ denote the indicator function and define $\mathcal{G}$ to be the set of all deterministic functions $g : \mathcal{X} \to \mathcal{A}$ such that $\mathbb{E}[Cost_a(L)I(g(X) = a)] = \tau_a$ for all $a \in \mathcal{A}\backslash\{0\}$ and some positive reals $\tau_a \leq \mathbb{E}[Cost_a(L)]$. Suppose*

*$(Y^1 - Y^0) \perp\!\!\!\perp \varepsilon|L$, $\mathbb{E}[Y^1 - Y^0|L] \in \mathbb{R}$ and $\varepsilon|L \sim \text{Uniform}[0,1]$. Let $h_a(L) = \mathbb{E}[Y^a - Y^0|L]/Cost_a(L)$ for all $a \in \mathcal{A}\backslash\{0\}$ and define $g^*$ such that*

$$g^*((L,\varepsilon)) = \begin{cases} \min\left\{\underset{a\in\mathcal{A}\backslash\{0\}:\mathcal{P}(a,L)}{\arg\max} h_a(L)\right\} & \text{if } \mathcal{P}(a,L) \text{ for some } a \in \mathcal{A}\backslash\{0\}, \\ 0 & \text{otherwise} \end{cases}$$

*for all $(L,\varepsilon) \in \mathcal{X}$ and where $\mathcal{P}(a,L)$ is true if and only if $h_a(L) > k_a \vee [h_a(L) = k_a \wedge \varepsilon < p]$. Then, (i) there exist $k_a \in \mathbb{R} \cup \{-\infty\}$ and $p_a \in [0,1]$ for $a \in \mathcal{A}\backslash\{0\}$ such that $g^* \in \mathcal{G}$ and (ii)*

$$g^* \in \underset{g\in\mathcal{G}}{\arg\max}\ \mathbb{E}[Y^{g(X)}].$$

*Counterexample.* Let $\mathcal{A} = \{0,1,2\}$ and suppose $L$ is binary with $\Pr(L=1) = 1/2$. Suppose also that $Cost_a(L) = 1$ and that $\tau_a = 1/4$ for all $a \in \mathcal{A}\backslash\{0\}$. Suppose further that

$$\mathbb{E}[Y^a|L] = \begin{cases} 0 & \text{if } a = 0, \\ 5 & \text{if } a = 1 \wedge L = 0, \\ 4 & \text{if } a = 1 \wedge L = 1, \\ 4 & \text{if } a = 2 \wedge L = 0, \\ 1 & \text{if } a = 2 \wedge L = 1, \end{cases}$$

$$\text{so that} \quad h_a(L) = \begin{cases} 5 & \text{if } a = 1 \wedge L = 0, \\ 4 & \text{if } a = 1 \wedge L = 1, \\ 4 & \text{if } a = 2 \wedge L = 0, \\ 1 & \text{if } a = 2 \wedge L = 1. \end{cases}$$

Suppose now that $g^* \in \mathcal{G}$. Then, $k_1 = 5$, $k_2 = 1$ and $p_1 = p_2 = 1/2$. Indeed, if $k_1 > 5$, then $\mathcal{P}(1,L)$ is false for all $L$ and, so, $\mathbb{E}[g^*(X) = 1] = 0 \neq \tau_1$. If $k_1 < 5$, then $\mathcal{P}(1,L)$ is true for all $L$ and $\mathbb{E}[g^*(X) = 1] = \mathbb{E}[g^*(X) = 1|L = 0]/2 + \mathbb{E}[g^*(X) = 1|L = 1]/2 = 1 \neq \tau_1$. If $k_1 = 5$, then $\mathcal{P}(1,L)$ is true if and only if $L = 0$ and $\varepsilon < p$, so $\mathbb{E}[g^*(X) = 1] = \Pr(L = 0, \varepsilon < p) = \Pr(L = 0)\Pr(\varepsilon < p) = p/2$ and $p/2 = \tau_1 = 1/4$ if and only if $p = 1/2$. Similar arguments establish that $k_2 = 1$ and $p_2 = 1/2$ if $g^* \in \mathcal{G}$.

Hence,

$$\begin{aligned} \mathbb{E}[Y^{g^*(X)}] &= \mathbb{E}[Y^0 + (Y^1 - Y^0)I(g^*(X) = 1) + (Y^2 - Y^0)I(g^*(X) = 2)] \\ &= \mathbb{E}[Y^0] + \mathbb{E}[Y^1 - Y^0|g^*(X) = 1]\tau_1 + \mathbb{E}[Y^2 - Y^0|g^*(X) = 2]\tau_2 \\ &= \mathbb{E}[Y^1 - Y^0|L = 0, \varepsilon < 1/2]\tau_1 + \mathbb{E}[Y^2 - Y^0|L = 1, \varepsilon < 1/2]\tau_2 \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\big[Y^1 - Y^0 \big| L = 0\big]\tau_1 + \mathbb{E}\big[Y^2 - Y^0 \big| L = 1\big]\tau_2 \\
&= 5/4 + 1/4 = 1.5.
\end{aligned}
$$

Now, define $\tilde{g} : \mathcal{X} \to \mathcal{A}$ such that

$$
\tilde{g}((L, \varepsilon)) = \begin{cases} 1 & \text{if } L = 1 \wedge \varepsilon < 1/2, \\ 2 & \text{if } L = 0 \wedge \varepsilon < 1/2, \\ 0 & \text{otherwise,} \end{cases}
$$

so that $\mathbb{E}[\tilde{g}(X) = 1] = \tau_1$ and $\mathbb{E}[\tilde{g}(X) = 2] = \tau_2$. But

$$
\begin{aligned}
\mathbb{E}\big[Y^{\tilde{g}(X)}\big] &= \mathbb{E}\big[Y^0\big] + \mathbb{E}\big[Y^1 - Y^0 \big| \tilde{g}(X) = 1\big]\tau_1 + \mathbb{E}\big[Y^2 - Y^0 \big| \tilde{g}(X) = 2\big]\tau_2 \\
&= \mathbb{E}\big[Y^1 - Y^0 \big| L = 1\big]\tau_1 + \mathbb{E}\big[Y^2 - Y^0 \big| L = 0\big]\tau_2 \\
&= 4/4 + 4/4 = 2.
\end{aligned}
$$

Hence, $\mathbb{E}\big[Y^{\tilde{g}(X)}\big] > \mathbb{E}\big[Y^{g^*(X)}\big]$ and $\tilde{g} \in \mathcal{G}$ and, so, $g^* \notin \operatorname*{arg\,max}_{g \in \mathcal{G}} \mathbb{E}\big[Y^{g(X)}\big]$. □

# 12

# Summary and general discussion

Epidemiology is a broad field of study with methods and concepts connecting all subfields (Lau et al., 2020). This thesis describes a study of epidemiological methods for answering questions about cause and effect in the presence of methodological obstacles, such as confounding, missing data or measurement error. In this chapter, a summary of our main findings is presented, along with a general discussion of this thesis in the light of the existing literature, with suggestions for future research.

## 12.1 Summary of findings

Methods for answering causal questions can be studied with the aim of learning about its workings, its performance under certain conditions. At a more meta-level, we can study how methods are being disseminated or implemented. Likewise, we can study, on the one hand, how and when a methodological obstacle may be overcome, and, on the other, how it is handled in applied research. In **chapter 2**, we questioned some of the current practice of how research at this meta-level is conducted, particularly where it concerns the initial phases of a systematic literature review. The standard approach of ignoring the text body in searching or screening articles might fail to retrieve all or a representative sample of the relevant literature, potentially leading to a false impression about the topic of enquiry. We found that for a number of methodological topics, a large portion of articles with a topic mention somewhere in the text did not contain a reference to the topic in text fields other than the body. The results do not conclusively show that ignoring text bodies does indeed lead to a false impression, but it should raise suspicion. Researchers might wish to consider including these text fields in their search and screening strategy.

In primary research, epidemiologists are often faced with multiple methodological obstacles simultaneously. There are concerns, however, that combinations of methods designed for different methodological obstacles have worse performance than might be expected from how they perform in isolation. In **chapters 3 and 4**, we critically reflected on a previous simulation study by Mitra and Reiter (2016), in which they compare two approaches to implementing propensity score matching after multiply imputing missing data. We found that the standard multiple imputation approach of carrying out analysis *within* multiply imputed datasets before pooling the results is generally to be preferred over their proposed approach of first pooling propensity scores *across* multiply imputed datasets before carrying out matching (or any other propensity score method) based on these pooled scores. Our results are in stark contrast to the results of Mitra and Reiter (2016) and we argued that this is largely due to their use of a misspecified imputation model that ignores the outcome variable.

Propensity score estimation is typically done by fitting a logistic regression. However, standard regression modelling software by default discards all incomplete records and does not offer propensity score estimates for subjects with missing data. Machine learning techniques such as classification and regression trees (CART) are appealing in part because some implementations allow for incomplete records to be incorporated in the tree fitting and provide propensity score estimates for all subjects. An important question to be answered is whether and when CART handles the missing data in a desirable way. In **chapter 5**, we argued that the automatic handling of missing data by CART is by no means a one-fits-all solution to the problem of missing covariate data for causal inferences based on propensity score methods. In a number of simulation studies, we actually found CART to be outperformed by standard, alternative methods to account for missing data. Different CART implementations handle missing data differently. In judging whether a given implementation is appropriate for the task at hand, some understanding of the ‘black-box nature’ of machine learning algorithms is therefore desirable.

In **chapter 6**, we considered missing outcome rather than missing covariate data. The chapter gives no new results but emphasises and illustrates that when baseline exchangeability is achieved through propensity score matching, bias might nonetheless result from restricting downstream analysis to the subset of individuals who have not dropped out of the study by the administrative study end. This equally applies to controlled trials with baseline randomisation, where exchangeability, achieved at baseline by design, is not guaranteed to uphold in the set of complete records that may be used for the analysis. Regression and inverse probability of censoring weighting were discussed as possible solutions.

Researchers can sometimes have a considerable influence over the extent of missingness. In studies on the effects of time-varying exposures, information of post-baseline covariates may help mitigate time-dependent confounding, but obtaining a record of the values that these variables take at each of potentially many time points can be costly and time-consuming. Reducing the frequency of measurements may enhance study feasibility, but it may also compromise study validity. In **chapter 7**, we illustrated by way of simulation the impact of choices regarding the frequency of measuring time-varying covariates. To handle missing values, we implemented the last-observation-carried-forward procedure (LOCF) under the implicit (and wrong) assumption that the participant characteristics remained constant in periods of no measurement. As expected, in our simulations, fixed-interval measurement resulted in bias consistent with residual confounding. We additionally showed that bias might arise in settings where decisions to measure are driven by observed values of the time-varying exposure, such as in the studies of Ali et al. (2016) and Souverein et al. (2016).

When variables take values that are different from what these values appear or are assumed to be, such as may be the case when we implement LOCF, we say that the variables are subject to measurement error. When the variables are categorical, we speak of misclassification, a special type of measurement error. In **chapter 8**, we focused on joint exposure-outcome misclassification and developed a method for this issue in the presence of confounding. Simulation studies showed favourable large sample performance. However, further research is needed to study the sensitivity of the proposed method and that of alternatives to violations of their assumptions.

Concerns about violations of assumptions are common in observational research on causal effects. In efforts to lessen these concerns, it has been suggested that so-called negative control variables are used (Lipsitch et al., 2010). Negative controls are variables that are known (or at least believed) to be causally unrelated to one or more of the variables of interest. The key idea is that observing an association that contradicts the belief in a causal null relation might alert the analyst to violations of assumptions. Negative controls have potential in bias detection as well as partial or complete bias correction in epidemiological research. In **chapter 9**, we sought to complement efforts to increase the more routine use of negative controls with a discussion about a selection of caveats. We argued that negative controls may lack both specificity and sensitivity to detect unmeasured confounding. We also reviewed existing methods to adjust for unmeasured confounding based on negative controls and examined the impact of assumption violations. Given the potentially large impact, it may sometimes be desirable to replace strong conditions for exact identification with weaker, easily verifiable conditions, even when these imply at most partial identification. Future research in this area may broaden the applicability of negative controls and in turn make them better suited for routine use in epidemiological practice. At present, however, the applicability of negative controls should be carefully judged on a case-by-case basis.

Case-control designs are an important tool in causal inference. In **chapter 10**, we argued that to facilitate understanding, it is useful to consider every case-control study as being nested within a cohort study. The case-control study then effectively becomes a cohort study with missingness governed by the control-sampling scheme. In the chapter, we gave an overview of how observational data obtained with case-control designs can be used to identify a number of causal estimands and, in doing so, recast historical case-control concepts, assumptions and principles in a modern and formal framework.

Finally, in **chapter 11**, we turned to precision medicine and considered the task of finding the optimal subgroup for treatment under certain cost or resource constraints. In practice, it is not uncommon for treatment assignment decisions to be based of prognostic scores. However, this approach does not guarantee optimal results (VanderWeele et al., 2019). As an alternative, one may attempt to evaluate all possible subgroups one by one, and choose the rule with the 'best' results. However, this is not feasible when there are many, potentially infinitely many subgroups to consider. VanderWeele et al. (2019) showed that the task can sometimes be considerably simplified by deriving treatment assignment rules that (1) guarantee optimality under some conditions and (2) take a simple form: assign treatment in a greedy fashion to all individuals with the next largest benefit (i.e., the largest difference in potential outcome means given covariates) or the next highest benefit–cost ratio (with cost being a positive function of baseline covariates) until the resource or cost constraint, respectively, is exceeded. The optimality of the rules however relies critically on the assumption that there are no tied conditional treatment effects or benefit-cost ratios between individuals. We extended their work by deriving rules that likewise have a simple form and which guarantee optimality under the same conditions, except that there need be no constraint on the presence of ties. An important insight that this chapter is meant provide is that in order to obtain some sense of optimally from allocating treatment, a contrast between counterfactual outcomes under different treatment options should be considered. Prognostic scores alone are not (generally) sufficient. The methodological obstacles that we encounter in causal inference, including confounding, missing data and measurement error, are therefore relevant in precision medicine too.

## 12.2 General discussion

The methodological aspects of causal inference form a broad topic and we addressed a variety of subtopics in this thesis. Apart from confounding, missing data, and measurement error, the reader may nonetheless recognise a number of recurrent features.

For example, Monte Carlo simulation was used in a number of chapters (e.g., **chapters 3-8**). This is a useful tool for obtaining empirical results (i.e., approximations) about the performance of statistical methods in certain scenarios as opposed to more general, analytic results (Morris et al., 2019). They are particularly appealing when the latter are difficult to obtain, or when the interest lies with illustrating a problem or method. However, they also have limitations. They provide at most approximations of statistical properties. Also, only a limited, finite number of scenarios can be considered and there is often the concern that the results generalise poorly to other scenarios.

Much of this thesis is built on the potential or counterfactual outcomes framework. In this work, like much of the literature, the terms ‘potential outcomes’ and ‘counterfactual outcomes’ are used interchangeably. Where they are considered distinct, generally the potential and counterfactual versions of a variable under the same hypothetical situation are still regarded as having the same values. However, variables are labeled as potential or counterfactual depending on whether they are seen as primitive or constructed from a collection of functions and background variables, respectively (Pearl, 2010). In some parts of this thesis (e.g. **chapter 5**), we explicitly took a constructivist approach, while in others (e.g., **chapter 10**), we did not. The adjective ‘potential’ further connotes a prospective view; either one of multiple versions of the outcome might become real-world before the choice among the corresponding mutually exclusive actions is made. By contrast, ‘counterfactual’ connotes a retrospective view; the choice among mutually exclusive actions is made and all but one version of the outcome is contrary-to-fact.

The notion of ‘counterfactual thinking’ is not used merely in epidemiology and has found its way in many branches of science, including physics (Robins et al., 2015). Its uptake and popularity in epidemiology, however, have given rise to much dispute among academics (Vandenbroucke et al., 2016; Krieger and Davey Smith, 2016; Broadbent et al., 2016; VanderWeele, 2016; VanderWeele et al., 2016; Schwartz et al., 2016; Daniel et al., 2016; Robins and Weissman, 2016;

Blakely et al., 2016). A central point of critique is that counterfactual thinking would delimit the meaning of causality by equating "causal claims with precise predictions about contrary-to-fact scenarios" (Vandenbroucke et al., 2016). A contrasting view is that the counterfactual framework considers a subset—not necessarily the entire set—of causal claims, namely those that can be phrased as statements about the consequences of hypothetical—possibly contrary-to-fact—actions (VanderWeele et al., 2016). Sometimes, the framework might admit non-actions (e.g., states) as potential causes but only when it is understood what actions are implied. The focus on this subset of causal claims is meant to guide decisions in the real world based on predictions of their consequences.

It should be noted that even after restricting to this subset of causal questions, some ambiguity about what the actions (interventions) and corresponding counterfactuals mean often remains. This issue relates to another point of debate: the well-definedness of interventions and counterfactuals. It is important to note that well-definedness of interventions is not the same as the interventions being elaborate. Telling a patient to follow a poorly detailed drug prescription or exercise programme, and advising social distancing against the spread of COVID-19 during a given press conference may well represent reasonably well-defined (point) interventions. They are not made less well-defined by the patient being unsure of how to interpret the drug prescription or exercise programme, or by the residents of a country not acting on the social distancing advise in a uniform way. Well-definedness of interventions relates to the lack of ambiguity of what the interventions mean, not about how they should be acted on. The requirement that interventions and counterfactuals are sufficiently well-defined, as noted in the introduction of this thesis, is that there is no ambiguity about the interventions or that the counterfactuals are invariant to the choice among the possible variations. Striving for well-definedness only serves to eliminate vagueness about the meaning of a causal effect.

Other critique relates to the assumptions that can be readily made explicit with counterfactual parlance (Schwartz et al., 2016), and the tools that are typically associated with or embedded in the counterfactual framework, such as directed acyclic graphs (DAGs) or single-world intervention graphs (sWIGs) (Richardson and Robins, 2013) with which some assumptions can be graphically encoded. However, that the assumption of, say, 'no interference' for a joint intervention on multiple individuals (i.e., 'one individual's treatment does not affect another's outcome') is often made (albeit often implicitly) or can be articulated with relative ease, does not mean that the counterfactual framework permits only causal inference under this assumption (Robins and Weissman, 2016). The development of a language rich enough to articulate a wider variety of causal questions and assumptions is an advance with positive effects on clarity of thought and ease of communication. The assumptions that are made explicit and least ambiguously articulated are inevitably often the ones that receive the most scrutiny and criticism. As Pearl et al. (2014) notes, “he who seeks licensing assumptions risks suspicions of attempting to endorse those assumptions. ... The more explicit the assumption, the more criticism it invites”. Methodological decisions (e.g., about which variables to ‘adjust’ for, or about the use of complete case analysis versus multiple imputation for missing data) often rely on structural assumptions about the data. There are often concerns that the DAGs encoding data structures are too simplistic. Robins (1999) argues that although the real world may well be more complex than is sometimes implied by a simple graph, “if we do not learn how to reason correctly in simple causal Gedankenexperiments ..., we have no chance of success in realistic situations.” Uncertainty about whether certain (identifiability) assumptions are met does not justify that potential assumption violations are ignored or rigour abandoned.

Like ‘counterfactuals’, ‘missing data’ and ‘measurement error’ are terms whose meaning is not always clear. For example, it is easy to conflate a given variable being inaccessible to the researcher (often encoded with ‘NA’) with the variable being accessible yet taking the value ‘missing’ or ‘NA’. For example, in an attempt to address confounding, one might wish to capture all information upon which a general practitioner (GP) bases his treatment decisions. The GP might fail to take a patient’s blood pressure, but this does not mean that the corresponding variable is truly missing. The GP cannot base decisions on what he did not observe and, so, the researcher might still have access to all variables that have informed the GP’s decision making. Similar comments apply to the notion of measurement error. Measurement error is a relative notion: in one context, systolic blood pressure plus some random term might be considered measurement error; in others, it is exactly what the researcher set out to measure.

### *Future perspectives*

Epidemiology continues to face both opportunities and challenges. The potential access to big data provides opportunities (e.g., for artificial intelligence and machine learning), but with increased use of data that are not collected for non-research purposes it is likely that methodological obstacles such as confounding, missing data and measurement error are becoming more prevalent or more severe. It is sometimes claimed that data collected for research purposes do not reflect daily practice. It is important to recognise, however, that, conversely, evidence that originates from daily practice does not necessarily provide valid evidence for daily practice. In the presence of difficult challenges, it is tempting to change one's inferential goals so that they become easier to achieve. However, this may leave the question that is of actual interest unanswered. If the interest is with a causal estimand, researchers should be explicit about this (Hernán, 2018).

Along with committing to a causal estimand, use of a causal roadmap may help avoid conflation of different parts of causal inference (Petersen and Van der Laan, 2014; Ahern, 2018). We believe that a distinction between identification and estimation is particularly useful as it means that the purely statistical issues of the latter can be put aside when concentrating on the former. At each step of the roadmap, there are areas for future methodological research.

For example, regarding missing data, emphasis is often placed on the classification of missingness as either being 'completely at random' (MCAR), 'at random' (MAR), or 'not at random' (NMAR), or on the recoverability of the entire joint distribution of a collection of variables. However, specific causal estimands might be identifiable even if the entire joint distribution cannot be recovered. For example, in case-control studies, the topic of **chapter 11**, certain causal effects may actually be identifiable from the observed data distribution while absolute risks are typically not.

When estimands are not identifiable, it may be possible to obtain partial identification bounds, which may preclude the estimand from taking, say, the null value of no causal effect. Partial identification is an interesting area for future research in part because it may inform sensitivity analyses.

Finally, rather than concentrating on methodological obstacles in isolation, we believe there may be value in considering multiple problems together (Van Smeden et al., 2021). After all, in applied research, epidemiologists often face multiple problems simultaneously and how they are best handled together is rarely obvious.

## References

Ahern, J. (2018): "Start with the "C-word," follow the roadmap for causal inference," *American Journal of Public Health*, 108, 621.

Ali, M. S., R. H. Groenwold, S. V. Belitser, P. C. Souverein, E. Martín, N. M. Gatto, C. Huerta, H. Gardarsdottir, K. C. Roes, A. W. Hoes, et al. (2016): "Methodological comparison of marginal structural model, time-varying cox regression, and propensity score methods: the example of antidepressant use and the risk of hip fracture," *Pharmacoepidemiology and drug safety*, 25, 114–121.

Blakely, T., J. Lynch, and R. Bentley (2016): “Commentary: DAGs and the restricted potential outcomes approach are tools, not theories of causation,” *International journal of epidemiology*, 45, 1835–1837.

Broadbent, A., J. P. Vandenbroucke, and N. Pearce (2016): “Response: formalism or pluralism? A reply to commentaries on ‘Causality and causal inference in epidemiology’,” *International Journal of Epidemiology*, 45, 1841–1851.

Daniel, R. M., B. L. De Stavola, and S. Vansteelandt (2016): “Commentary: The formal approach to quantitative causal inference in epidemiology: misguided or misrepresented?” *International journal of epidemiology*, 45, 1817–1829.

Hernán, M. A. (2018): “The C-word: scientific euphemisms do not improve causal inference from observational data,” *American journal of public health*, 108, 616–619.

Krieger, N. and G. Davey Smith (2016): “The tale wagged by the dag: broadening the scope of causal inference and explanation for epidemiology,” *International journal of epidemiology*, 45, 1787–1808.

Lau, B., P. Duggal, S. Ehrhardt, H. Armenian, C. C. Branas, G. A. Colditz, M. P. Fox, S. E. Hawes, J. He, A. Hofman, et al. (2020): “Perspectives on the future of epidemiology: a framework for training,” *American journal of epidemiology*, 189, 634–639.

Lipsitch, M., E. T. Tchetgen, and T. Cohen (2010): “Negative controls: a tool for detecting confounding and bias in observational studies,” *Epidemiology (Cambridge, Mass.)*, 21, 383.

Mitra, R. and J. P. Reiter (2016): “A comparison of two methods of estimating propensity scores after multiple imputation,” *Statistical methods in medical research*, 25, 188–204.

Morris, T. P., I. R. White, and M. J. Crowther (2019): “Using simulation studies to evaluate statistical methods,” *Statistics in medicine*, 38, 2074–2102.

Pearl, J. (2010): “On the consistency rule in causal inference: Axiom, definition, assumption, or theorem?” *Epidemiology*, 21.

Pearl, J., E. Bareinboim, et al. (2014): “External validity: From do-calculus to transportability across populations,” *Statistical Science*, 29, 579–595.

Petersen, M. L. and M. J. Van der Laan (2014): “Causal models and learning from data: integrating causal modeling and statistical estimation,” *Epidemiology (Cambridge, Mass.)*, 25, 418.

Richardson, T. S. and J. M. Robins (2013): “Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality,” *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128, 2013.

Robins, J. M. (1999): “[Choice as an alternative to control in observational studies]: comment,” *Statistical Science*, 14, 281–293.

Robins, J. M., T. J. VanderWeele, and R. D. Gill (2015): “A proof of bell’s inequality in quantum mechanics using causal interactions,” *Scandinavian Journal of Statistics*, 42, 329–335.

Robins, J. M. and M. B. Weissman (2016): “Commentary: counterfactual causation and streetlamps: what is to be done?” *International journal of epidemiology*, 45, 1830–1835.

Schwartz, S., S. J. Prins, U. B. Campbell, and N. M. Gatto (2016): “Is the “well-defined intervention assumption” politically conservative?” *Social science & medicine (1982)*, 166, 254.

Smeden, M., van, B. B. L. Penning de Vries, L. Nab, and R. H. H. Groenwold (2021): “Approaches to addressing missing values, measurement error, and confounding in epidemiologic studies” *Journal of Clinical Epidemiology*, 131, 89–100.

Souverein, P. C., V. Abbing-Karahagopian, E. Martin, C. Huerta, F. de Abajo, H. G. Leufkens, G. Candore, Y. Alvarez, J. Slattery, M. Miret, et al. (2016): “Understanding inconsistency in the results from observational pharmacoepidemiological studies: the case of antidepressant use and risk of hip/femur fractures,” *pharmacoepidemiology and drug safety*, 25, 88–102.

Vandenbroucke, J. P., A. Broadbent, and N. Pearce (2016): “Causality and causal inference in epidemiology: the need for a pluralistic approach,” *International journal of epidemiology*, 45, 1776–1786.

VanderWeele, T. J. (2016): “Commentary: on causes, causal inference, and potential outcomes,” *International journal of epidemiology*, 45, 1809–1816.

VanderWeele, T. J., M. A. Hernán, E. J. Tchetgen Tchetgen, and J. M. Robins (2016): "Re: Causality and causal inference in epidemiology: the need for a pluralistic approach," *International journal of epidemiology*, 45, 2199–2200.

VanderWeele, T. J., A. R. Luedtke, M. J. van der Laan, and R. C. Kessler (2019): "Selecting optimal subgroups for treatment using many covariates," *Epidemiology (Cambridge, Mass.)*, 30, 334.

# 13

# Dutch summary | Nederlandse samenvatting

Epidemiologie, letterlijk vertaald vanuit het Grieks als de studie (logos) van wat zich boven (epi) het volk (demos) begeeft, is wetenschappelijk onderzoek dat wordt gekenmerkt door vragen over het vóórkomen van gezondheids- of ziektetoestanden. Er wordt onderscheid gemaakt tussen theoretische en toegepaste epidemiologie. Waar de eerste tak zich richt op methoden om eerdergenoemde vragen te beantwoorden, staan bij de tweede tak juist de antwoorden op deze vragen centraal.

Een onderverdeling kan ook worden gemaakt op basis van of de interesse ligt bij causaliteit, of specifiek bij zogenoemde wat-alsvragen. Daarbij staat centraal de mate waarin de gevolgen van hypothetische, mogelijk contrafactische, handelingen van elkaar verschillen. In epidemiologisch onderzoek staan de handelingen bijvoorbeeld voor verschillende behandelopties en slaan de gevolgen op een aspect van gezondheid of welzijn. Kennis over de gevolgen van verschillende behandelopties is nuttig bij het zoeken naar die optie die het 'beste' is, dat wil zeggen de optie met de gevolgen die het meest gunstig worden geacht.

In veel gevallen ligt dit soort kennis echter niet voor het oprapen. We kunnen immers niet in de toekomst kijken, laat staan dat we de toekomsten die voortvloeien uit verschillende, elkaar-uitsluitende opties met elkaar kunnen vergelijken om vervolgens de beste optie te kiezen. In deze gevallen keert men zich daarom vaak tot bronnen van vergelijkingsmateriaal en constateert men wat er reeds gebeurde in gevallen waarbij een bepaalde optie was gekozen en wat er gebeurde in gevallen waarbij een andere optie was gekozen. Hoe vertalen deze constateringen zich naar situaties waar nog een keuze moet worden gemaakt? Bij de vertaalslag komen een aantal obstakels kijken.

In dit proefschrift wordt verslag gedaan van studies over epidemiologische methoden voor het beantwoorden van wat-alsvragen in aanwezigheid van methodologische obstakels, zoals zogenoemde verstoring ('confounding'), ontbrekende gegevens, en meetfouten. Hieronder volgt een samenvatting van de voornaamste bevindingen. Enige voorkennis over epidemiologische methoden wordt verondersteld.

In het merendeel van de studies uit dit proefschrift werd onderzocht hoe methoden presteren in bepaalde situaties, 'hoe goed ze werken'. Onderzoek naar methoden kan echter ook gaan over hoe en wanneer deze in de praktijk worden ingezet. In **hoofdstuk 2** werpen we een kritische blik op hoe onderzoek op dit metaniveau in de praktijk wordt uitgevoerd. We stellen dat het negeren van de hoofdtekst bij het zoeken naar en screenen van artikelen (een standaard aanpak in systematisch literatuuronderzoek) mogelijk niet volstaat om alle of een representatieve steekproef van de relevante literatuur te verzamelen, hetgeen mogelijk leidt tot een vertekend beeld van het onderzoeksonderwerp. Voor een aantal methodologische onderwerpen vonden we dat in veel van de artikelen waarin het onderwerp ergens werd genoemd, geen verwijzing werd gegeven naar het onderwerp in tekstvelden anders dan de hoofdtekst. Onderzoekers zouden mogelijk willen overwegen om ook de hoofdtekst mee te nemen in de eerste fasen van systematisch literatuur onderzoek.

Epidemiologen hebben vaak om te gaan met meerdere obstakels. Methoden voor de verschillende obstakels kunnen soms eenvoudig worden gecombineerd. Echter, hoe goed de combinaties werken wijkt mogelijk af van hoe de methoden werken voor elk van de obstakels apart. In **hoofdstukken 3 en 4** reflecteren we op een eerder uitgevoerde simulatiestudie waarin twee combinaties van een methode voor ontbrekende gegevens, meervoudige imputatie, en een methode voor verstoring, 'propensity score matching', werden vergeleken. Kenmerkend voor meervoudige imputatie is dat er meerdere complete datasets worden gevormd. Een methode als propensity score matching kan op elk van de datasets afzonderlijk worden uitgevoerd. Een eerste stap, het schatten van zogenoemde propensity scores, zou dan voor elke dataset en elk onderzoekspersoon een geschatte propensity score opleveren. Volgens de standaard meervoudige-imputatie-aanpak (optie 1) wordt voor ieder behandeld (of blootgestelde) persoon vervolgens gezocht naar een onbehandelde persoon uit dezelfde geïmputeerde dataset met een soortgelijke schatting van de propensity score ('matchen', of koppelen). Een alternatief (optie 2) is om te matchen op basis van het gemiddelde van de propensity scores van een individu, genomen over alle geïmputeerde datasets. Beide opties leiden tot een verzameling van koppels en uiteindelijk een effectschatting voor elk van de geïmputeerde datasets. De laatste stap is het middelen van de effectschattingen. We concludeerden dat optie 1 in het algemeen te verkiezen is boven optie 2. Deze conclusie staat in schril contrast met die van de eerder uitgevoerde simulatiestudie. Dit werd deels verklaard door het gebruik in de eerdere studie van een ongeschikt imputatiemodel waarin de uitkomstvariabele wordt genegeerd.

Het is gangbaar propensity scores te schatten door middel van een logistische regressie. Standaard software laat echter alle individuën met ontbrekende gegevens buiten beschouwing. Machine learningtechnieken zoals 'classification and regression trees' (CART) zijn aantrekkelijk deels omdat sommige varianten alle individuën in een dataset, zo ook die met ontbrekende gegevens, meenemen. Een belangrijke vraag is echter of en wanneer CART dat doet op een wenselijke manier. In **hoofdstuk 5** beargumenteerden we dat het automatisch meenemen van ontbrekende gegevens geen universele oplossing is voor het probleem van ontbrekende gegevens bij causaal onderzoek op basis van propensity scores. In een aantal simulatiestudies was CART ondergeschikt aan standaard alternatieven. Verschillende CART-varianten gaan anders om met ontbrekende gegevens. Om te bepalen welke aanpak geschikt is in een bepaalde setting achten we het daarom wenselijk om de 'black box' van machine learningalgoritmes enigszins te doorgronden.

In **hoofdstuk 6** staan ontbrekende waarden van de uitkomstvariabele centraal. Het hoofdstuk beschrijft geen nieuwe resultaten maar illustreert dat de zogenoemde 'uitwisselbaarheid' tussen behandelgroepen aan het begin van een waarnemingsperiode (bijvoorbeeld verkregen door propensity score matching) kan worden gecompromitteerd door te conditioneren op het geobserveerd hebben van uitkomstwaarden. Dit geldt evenzo voor gerandomiseerd onderzoek waar uitwisselbaarheid aan het begin van de waarnemingsperiode door randomisatie wordt verkregen. Regressieanalyse en inversekansweging worden in het hoofdstuk genoemd als mogelijke oplossingen.

Onderzoekers hebben soms behoorlijke invloed op de mate waarin gegevens ontbreken. In studies naar de effecten van tijdsafhankelijke blootstellingen kan informatie over 'post-baseline' variabelen nuttig zijn om te corrigeren voor tijdsafhankelijke verstoring. Echter, het verzamelen van voldoende gedetailleerde informatie kan kostbaar en tijdrovend zijn. Het beperken van de meetfrequentie kan de uitvoerbaarheid van een studie ten goede komen, maar tegelijkertijd kan het de validiteit van het onderzoek in gedrang brengen. In **hoofdstuk 7** worden simulatiestudies beschreven die de impact van keuzes ten aanzien van meetfrequenties van tijdsafhankelijke variabelen illustreren. Hierbij werd het last-observation-carried-forwardprincipe (LOCF) gebruikt onder de impliciete (en verkeerde) aanname dat patiëntkarakteristieken onveranderd bleven in periods waarin niet gemeten werd. Zoals in eerder onderzoek al is laten zien gingen meetstrategiën met constante intervallen tussen meetmomenten gepaard met systematische verschillen tussen effectschattingen en de effecten die geschat werden. Dit hoofdstuk illustreert daarnaast dat systematische verschillen ook ontstaan wanneer meetmomenten worden geselecteerd op basis van eerder-geobserveerde gegevens.

Wanneer de eigenlijke waarden van variabelen anders zijn dan wat wordt geregistreerd of aangenomen (zoals bijvoorbeeld door LOCF), dan wordt er gesproken van meetfouten. Wanneer het categorische variabelen betreft wordt ook de term misclassificatie gebruikt. In **hoofdstuk 8** ligt de focus op gezamenlijke blootstellings- en uitkomstmisclassificatie en wordt een nieuwe methode beschreven om voor zowel deze vorm van meetfouten alsook voor confounding te corrigeren. Simulatiestudies brachten gunstige eigenschappen voor grote steekproeven aan het licht. Verder onderzoek is echter nodig om in kaart te brengen hoe goed de methode presteert wanneer aannames worden geschonden.

Zorgen over schendingen van aannames zijn veelvoorkomend in observationeel onderzoek naar oorzakelijkheid. Om de zorgen te verminderen is voorgesteld om gebruik te van zogenoemde negative controls, variabelen waarvan bekend is (of wordt aangenomen) dat ze geen oorzakelijk verband hebben met de primaire blootstelling of uitkomstvariabele. Het onderliggende idee is dat een waargenomen associatie die tegenstrijdig is met deze negative controlaanname een indicatie is voor een aannameschending. In **hoofdstuk 9** vullen we aanmoedigingen voor routinematig gebruik van negative controls aan met een beschouwing van limitaties. We beargumenteren dat negative controls soms verkeerde signalen geeft over de aan- of afwezigheid van verstoring. Het hoofdstuk geeft ook een beschouwing van bestaande negative controlemethoden voor correctie van verstoring en illustreert ook de gevoeligheid van deze methoden voor aannameschendingen. Gezien de mogelijk sterke impact van aannameschendingen achten we het soms wenselijk om de sterke aannames voor exacte identificatie van causale effecten te vervangen met zwakkere en makkelijker te verifiëren aannames, zelfs wanneer deze slechts gedeeltelijke identificatie garanderen. Verder onderzoek hiernaar kan mogelijk het nut van negative controls als standaard methode in epidemiologisch onderzoek verder motiveren. Vooralsnog zal de geschiktheid van negative controls zorgvuldig per geval moeten worden beoordeeld en terughoudendheid worden betracht.

Patiënt-controleonderzoek vormt een belangrijk gereedschap voor onderzoek naar causaliteit. In **hoofdstuk 10** wordt een patiënt-controlestudie conceptueel voorgesteld als een cohortstudie met ontbrekende gegevens, waarbij het ontbreken van gegevens wordt bepaald door de controleselectieprocedure. Vanuit dit perspectief geeft het hoofdstuk een overzicht van hoe bepaalde causale effecten kunnen worden geïdentificeerd met patiëntcontroleonderzoek. Hierbij worden traditionele concepten, aannames en principes geplaatst in een modern en formeel raamwerk.

Tot slot richten we ons in **hoofdstuk 11** op persoonsgerichte (in plaats van groepsgerichte) gezondheidszorg en specifiek tot de taak om zo optimaal mogelijk een mogelijk beperkte hoeveelheid behandelingen te verdelen over een groep patiënten. In de praktijk is het niet ongebruikelijk om behandelkeuzes te baseren op prognostische scores. Deze aanpak is echter niet altijd optimaal. Als alternatief kan men een voor een alle mogelijke behandelstrategieën afgaan om vervolgens de 'beste' te kiezen. Praktisch gezien is dit echter niet uitvoerbaar wanneer er veel (mogelijk oneindig veel) strategieën te bedenken zijn. In eerder theoretisch onderzoek werd aangetoond dat de taak soms aanzienlijk kan worden vereenvoudigd door algemene behandelregels af te leiden die (1) een zekere optimaliteit garanderen onder bepaalde aannames en (2) een heel simpele vorm aannemen: wijs behandeling toe aan de individuen met het grootste behandelvoordeel (dat wil zeggen, het grootste verschil in gemiddelde contrafactische uitkomsten gegeven baseline patiëntkarakteristieken tussen wel en niet behandelen) of de hoogste voordeel-kostenquotiënt (waarbij de kosten als positieve functie van de baseline patiëntkarakteristieken worden gezien) en aan een zo'n groot mogelijk deel van alle patiënten zodat een vooropgestelde grens op het aantal behandelingen dan wel de totale kosten niet wordt overschreden. De optimaliteit van de regels berust op de aanname dat patiënten met verschillende baseline karakteristieken verschillende behandelvoordelen of voordeel-kostenquotiënten hebben. In het hoofdstuk wordt een simpele aanpassing van de regels voorgesteld die er voor zorgt dat de optimaliteit gewaarborgd blijft zelfs wanneer de aanname wordt geschonden. Een belangrijk inzicht dat het hoofdstuk geeft is dat om een zekere optimaliteit te verkrijgen antwoorden op wat-alsvragen soms nodig zijn. Prognostische scores zijn op zichzelf niet voldoende. De methodologische obstakels in onderzoek naar causaliteit, zoals verstoring, ontbrekende gegevens en meetfouten, zijn daarom ook relevant in de persoonsgerichte gezondheidszorg.

# Acknowledgements | Dankwoord

Ten opzichte van andere proefschriften houd ik mijn dankwoord beknopt en ik vertrouw erop dat de lezer dit niet opvat als teken van ondankbaarheid. Ik noem een aantal van hen wier betrokkenheid bij mijn promotietraject ik zeer heb gewaardeerd: Rolf Groenwold, die het op zich nam om mij als masterstudent in Utrecht en later als promovendus in Utrecht en Leiden te begeleiden en met wie ik vaak plezierig van gedachten heb mogen wisselen; iedereen, en in het bijzonder Kim Luijken, Linda Nab en mijn co-promotor Maarten van Smeden, met wie ik gelijktijdig deel heb uitgemaakt van Rolf's onderzoeksgroep; professor Miguel Hernán en diens collega's die mij als gastonderzoeker verwelkomden en een kijkje lieten nemen in Boston en Harvard University; eens C7-107-kamergenoot Ype de Jong en andere collega's van de afdeling Klinische Epidemiologie waarmee ik het goed kon vinden; en bovenal pap, mam, mijn broers Joost en Tom en ook mijn vriendin Paula, die altijd het beste met mij voor hebben.

# Curriculum vitae

Bas Penning de Vries was born on May 4th, 1992, in Hilversum, the Netherlands. In 2010, he graduated from secondary school (A. Roland Holst College, atheneum) and enrolled for an integrated master's degree in chiropractic at the Anglo-European College of Chiropractic (AECC) in Bournemouth, England. Five years later, he graduated from the AECC with distinction and entered a master's programme in epidemiology with a specialisation in medical statistics at Utrecht University. As part of an internship at the Julius Center for Health Sciences and Primary Care of University Medical Center (UMC) Utrecht, Bas began doing methodological research under supervision of prof.dr. Rolf Groenwold. After obtaining his master's degree in epidemiology (*cum laude*) in 2017, Bas continued his work with prof.dr Rolf Groenwold as a PhD candidate, first at UMC Utrecht and later at the Department of Clinical Epidemiology of Leiden University Medical Center.

# List of publications

Penning de Vries, B. B. L., M. van Smeden, F. R. Rosendaal, R. H. H. Groenwold (2020): "A comparison between full text mining and searching in title, abstract and keywords for systematic reviews of epidemiological practice," *Journal of Clinical Epidemiology*, 121, 55–61.

Penning de Vries, B. B. L., R. H. H. Groenwold (2016): "Comments on propensity score matching following multiple imputation," *Statistical Methods in Medical Research*, 25(6), 3066–3068.

Penning de Vries, B. B. L., R. H. H. Groenwold (2017): "A comparison of two approaches to implementing propensity score methods following multiple imputation," *Epidemiology, Biostatistics and Public Health*, 14(4), e12630.

Penning de Vries, B. B. L., M. van Smeden, R. H. H. Groenwold (2018): "Propensity score estimation using classification and regression trees in the presence of missing covariate data," *Epidemiologic Methods*, 7(1).

Penning de Vries, B. B. L., R. H. H. Groenwold (2018): "Cautionary note: propensity score matching does not account for bias due to censoring," *Nephrology, Dialysis and Transplantation*, 33(6): 914–916.

Penning de Vries, B. B. L., R. H. H. Groenwold (2021): "Bias of time-varying exposure effects due to time-varying covariate measurement strategies," *Pharmacoepidemiology and Drug Safety*, 1–6.

Penning de Vries, B. B. L., M. van Smeden, R. H. H. Groenwold (2021): "A weighting method for simultaneous adjustment for confounding and joint exposure-outcome misclassifications," *Statistical Methods in Medical Research*, 30(2): 473–487.

Penning de Vries, B. B. L., R. H. H. Groenwold (2021): “Negative controls: concepts and caveats”. [Submitted]

Penning de Vries, B. B. L., R. H. H. Groenwold (2021): “Identification of causal effects in case-control studies,” *BMC Medical Research Methodology*. [Accepted for publication]

Penning de Vries, B. B. L., R. H. H. Groenwold, A. Luedtke (2021): “On selecting optimal subgroups for treatment using many covariates,” *Epidemiology*, 31(4): e33–e34.