Exploring Objective Causal Inference in Case-Noncase Studies under the Rubin Causal Model

Exploring Objective Causal Inference in Case-Noncase Studies under the Rubin Causal Model

A dissertation presented
by

Nikola Andric

to
The Department of Statistics
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of

Statistics

Harvard University
Cambridge, Massachusetts
April 2015

Dissertation Advisor: Professor Donald B. Rubin **Nikola Andric**

Exploring Objective Causal Inference in Case-Noncase Studies under the Rubin Causal Model

Abstract

Case-noncase studies, also known as case-control studies, are ubiquitous in epidemiology, where a common goal is to estimate the effect of an exposure on an outcome of interest. In many areas of application, such as policy-informing drug utilization research, this effect is inherently causal. Although logistic regression, the predominant method for analysis of case-noncase data, and other traditional methodologies, may provide associative insights, they are generally inappropriate for causal conclusions. As such, they fail to address the very essence of many epidemiological investigations that employ them. In addition, these methodologies do not allow for outcome-free design (Rubin, 2007) of case-noncase data, which compromises the objectivity of resulting inferences.

This thesis is directed at exploring what can be done to preserve objectivity in the causal analysis of case-noncase study data. It is structured as follows.

In Chapter 1 we introduce a formal framework for studying causal effects from case-noncase data, which builds upon the well-established Rubin Causal Model for prospective studies.

In Chapter 2 we propose a two-party, three-step methodology — PrepDA — for objective causal inference with case-noncase data. We illustrate the application of our methodology in a simple non-trivial setting. Its operating characteristics are investigated via simulation, and compared to those of logistic and probit regression.

Chapter 3 focuses on the re-analysis of a subset of data from a published article, Karkouti et al. (2006). We investigate whether PrepDA and logistic regression, when applied to case-noncase data, can generate estimates that are concordant with those from the causal analysis of prospectively collected data. We introduce tools for covariate balance assessment across multiple imputed datasets. We explore the potential for analyst bias with logistic regression, when said method is used to analyze case-noncase data.

In Chapter 4 we discuss our technology's advantages over, and drawbacks as compared to, traditional approaches.

Contents

Acknowledgments

My most sincere thanks go to my advisor, Donald B. Rubin. I thank him for his guidance, encouragement and support during the development of this work. I also would like to thank D. James Greiner for his mentorship and support. My work as consulting expert in litigation, which D. Greiner supervised, influenced many of the ideas behind this thesis. I additionally am indebted to Mary Beth Landrum for her valuable feedback on my research.

I acknowledge faculty for their mentorship.

I thank my fellow graduate students, especially Viviana García-Horton, David Allan Watson, and Arman Sabbaghi, for their friendship, help, and the many discussions that enhanced my research.

Special thanks are due to Dr. Keyvan Karkouti, who provided me the data for Chapter 3 of this work.

I would like to express my gratitude to my family and friends for their continual encouragement and support.

Funding for this work was supported by NSERC (Natural Sciences and Engineering Research Council of Canada) scholarships PGS M and PGS D, and the FQRNT (Fonds Québécois de la Recherche sur la Nature et les Technologies) master's research scholarship.

To my family.

Chapter 1

Framework for causal inference in case-noncase studies

1.1 Introduction

1.1.1 The case-noncase study

The case-noncase study, also known as case-control study, is ubiquitous in epidemiology and biostatistics for screening factors suspected to be associated with rare diseases, and for quantifying how disease risk varies with said factors. It is, in fact, uniquely suited to the study of rare diseases, for which prospective cohort studies are often impractical due to prohibitive costs and restrictive logistics. A cohort study, for instance, may require follow-up of a sizable study population — possibly over an extensive period of time — to insure data collection on an adequate number of diseased units. In contrast, the case-noncase study generally allows for a faster and less costly investigation via selection of study subjects based on the outcome of interest. That is to say, the study starts *after* outcomes have been realized (and possibly observed). As such, the case-noncase design is retrospective and non-randomized (randomization cannot be used to assign units to particular treatments or regimes).

In its simplest form, the case-noncase study examines the relationship between a single pre-specified binary treatment and a binary outcome. For example, in policy-informing drug utilization research, the study might concern the effect of taking or not taking a particular medication (the treatment) on the subsequent occurrence or non-occurrence of disease (the outcome). In case-noncase jargon, individuals with a particular disease or condition are called 'cases', whereas those without the disease or condition are called 'noncases'[1]. After all outcomes have been realized, cases and noncases are sampled with differing probabilities. Cases are oversampled as a means to get around the issue of outcome rarity. We consider the scenario in which all cases have been sampled. This can be accomplished, say, through data collection via surveillance programs and registries. Noncases, on the other hand, are commonly selected via simple random sampling from an underlying cohort, whose existence is guaranteed under the setup of a case-cohort design[2] (Prentice, 1986), or by means of retrospective matching (Holland and Rubin, 1988). In this thesis we do not address studies of the latter type because retrospective matching is, in our opinion, an oft-misleading and generally inadequate method for pretreatment bias reduction in causal effect estimation (see Section 1.2.9). Instead, we focus on studies of the former type.

[1] Our use of the 'noncase' terminology intends to avoid confusion between the meanings of *control* as in non-occurrence of disease, and *control* as in non-exposure to active treatment.

[2] In a case-cohort design, a subcohort of noncases is randomly selected from a well-defined cohort.

Henceforth, we assume as research goal the quantification of the causal effect of a possible disease-causing treatment on an outcome of interest.

1.1.2 A brief literature review of the statistical analysis of case-noncase data

According to Breslow and Day (1980), an implementation of the case-noncase study was first reported in Lane-Claypon (1926). It was not until two decades later, however, that research articles on methodology and the statistical analysis of case-noncase data surfaced. Of particular importance were the classical Cornfield (1951) and Mantel and Haenszel (1959) papers. Cornfield (1951) showed that it was possible to estimate a relative risk from case-noncase data. Mantel and Haenszel (1959) introduced the χ^2 measure of statistical significance and a pooled estimator of relative risk. Mantel and Haenszel (1959) also discussed the role and limitations of the case-noncase design, and emphasized its relationship to cohort studies. Cornfield's early work gave way to a series of notable papers on the estimation of (log-)odds ratios (see, e.g., Woolf, 1955; Haldane, 1955; Gart, 1966). The 1970s brought generalizations of relative risk and odds ratio estimates to case-noncase designs with retrospective matching (e.g., Miettinen's (1970) estimation of relative risk from individually matched case-noncase studies).

An important subsequent development was the demonstration that logistic regression can be applied to the (associative) analysis of case-noncase data (Anderson, 1972, 1973; Mantel, 1973; Seigel and Greenhouse, 1973; Prentice and Pyke, 1979). Prentice and Pyke (1979), in particular, demonstrated that the "[odds ratios'] asymptotic variance matrices may be obtained by applying the original logistic regression model to the case-control study as if the data had been obtained in a prospective study". Logistic regression has since become the predominant method for the analysis of case-noncase study data. Gefeller et al.'s (1998) literature survey revealed a dramatic increase in its use over the 1955-1994 period: its rate of implementation rose from 18.4% in 1955, to a staggering 87.2% in 1994.

Among recent developments is Rose and van der Laan's (2009b; 2009a) work on nonparametric estimation of marginal causal effects from (retrospectively matched) case-noncase data via weighted targeted maximum likelihood estimation.

1.1.3 Causal inference under the Rubin Causal Model

The statistical framework for causal inference based on the idea of potential outcomes was first proposed by Neyman (1923) for randomized experiments. It was extended by Rubin (1974, 1977, 1978a) to non-randomized studies under a formal structure now commonly called *Rubin's Causal Model* (RCM, see Holland, 1986). The RCM consists of three components. The first is a set of fundamental notions: unit, treatment, covariate and potential outcomes. The second is the concept of an assignment mechanism. The third, and optional, component is a Bayesian model. We briefly outline each of these components below. The reader is referred to Imbens and Rubin (2015) for a comprehensive reference on the topic of causal inference in statistics, social sciences, and biomedical sciences under the RCM.

The first part of the RCM introduces four fundamental notions: unit, treatment, covariate and potential outcomes. A *unit* is a person or physical object at a particular point in time. Suppose that each unit in a study is subject to two *treatments*, or interventions (e.g., an active treatment and a control treatment) whose effects we wish to assess. For each of these units, we define the two *potential outcomes* (assuming SUTVA, see Holland, 1986) as the outcome that would be realized (an possibly observed) under the control treatment, and the outcome that would be realized (and possibly observed) under active treatment. Note that it is possible for outcomes to be realized but not observed. This can occur, for example, in a case-noncase study: a unit's outcome, although realized, will not be observed by the investigator if the unit under consideration is not selected into the case-noncase sample (see Sections 1.2.5 and 1.2.6 for further discussion). Because at most one of the treatments can be applied at any given time to any given unit, only the potential outcome corresponding to said applied treatment is realized, and hence observable. The other, i.e. the outcome that would have been realized had the alternative treatment been applied, is missing (Fundamental Problem of Causal Inference Holland, 1986). Accordingly, causal inference, or the inference of *causal effects* — which in turn are defined as the comparison of units' potential outcomes under the two treatments — can be formulated as a missing data problem. In tackling this problem, it is desirable to compare units in the active treatment group to those units from the control treatment group who share similar background characteristics. As such, causal inference methods generally take into account units' *covariates*, which are defined to be background characteristics that could not have been affected by treatment assignment.

The second component, the *assignment mechanism*, gives the probability of being assigned to the active treatment for each unit in the study as a function of units' covariates and, possibly, units' potential outcomes. This mechanism plays a central role in causal inference as it explains the occurrence of missing potential outcome data. What's more, the assignment mechanism enables researchers to understand, formulate and explicitly state any assumptions (e.g., unconfoundedness of the assignment mechanism, see Rubin, 1975; Imbens and Rubin, 2015) made in reaching causal conclusions.

The third, and optional, component of the RCM is a Bayesian probability model on the *science*, which is understood to be the triplet consisting of covariates, the potential outcome under active treatment, and the potential outcome under control treatment, for all units in the population. (The science, as such, is the object of causal inference.) The probability model is generally used by analysts to either (a) infer relevant super-population parameters, or (b) impute missing potential outcome data via posterior predictive sampling for purposes of finite-population inference.

1.1.4 Outcome-free design for objective causal inference

According to Rubin (2007), "typically in order to get a drug approved, US Food and Drug Administration (FDA) requires carefully specified randomized designs and carefully specified primary analyses and secondary supporting analyses, and often the data collection and first pass analyses are carried out by a [sic] agent independent from the organization trying to get approval for the drug. There is thus tremendous pressure to live with the answers that come from the pre-specified design and analyses." This modus operandi ensures objectivity in the investigation of causal effects from randomized experiments.

In contrast, “observational studies are generally fraught with problems that compromise any claim for objectivity of the resulting causal inferences” (Rubin, 2008). In regression adjustment, for instance, study design is not separated from outcome analysis. For this reason, it is both possible and tempting for researchers “to fish for a certain result, [by] fitting several models until the desired or expected answer appears” (Pattanayak et al., 2011). Chapter 3 explores this idea in the context of case-noncase data analysis.

Accordingly, Rubin (2007) advocates that “observational studies can and should be designed to approximate randomized experiments as closely as possible... [These studies] should be designed using only background information to create subgroups of similar treated and control units, where similar here refers to their distributions of background variables. Of great importance, this activity should be conducted without any access to any outcome data, thereby assuring the objectivity of the design.” By ‘design’, Rubin means all contemplation, collection, organization, and analysis of data that takes place prior to seeing any outcome data.

We henceforth define ‘objective’ causal inference as that whose design phase is ‘blinded’ to outcome data. Although ‘blinded causal inference’ might, as such, be more fitting terminology for the topic discussed in this thesis, we proceed with the term ‘objective’ for purposes of consistency with Rubin (2007).

1.1.5 Case-noncase study applications, and the need for objective and causal inference

Applications of case-noncase studies span a wide and diverse range of fields, from public health policy, to drug utilization research, to litigation support. For example, Centers for Disease Control and Prevention (CDC) investigated in 2006 the association between Fusarium keratitis, a rare and dangerous fungal infection of the cornea, and use of Bausch & Lomb's ReNu with MoistureLoc® contact lens solution. The agency concluded an increased risk for Fusarium keratitis associated with use of the solution (Barry et al., 2006). These findings had regulatory, market, and litigative implications. Soon after the agency posted its report, FDA recommended that "consumers... stop using ReNu with MoistureLoc® immediately" (Schultz, 2006). A month later, Bausch & Lomb announced its decision "to voluntarily recall and permanently remove this contact lens solution from the worldwide market" (Barry et al., 2006). It is reported that between 2008 and 2009, Baush & Lomb has "settled nearly 600 fungal-infection lawsuits" (USA Today, 2009). Another example is Raz et al. (2014), a topical study that found a positive association between maternal exposure to particulate matter air pollution, and odds of autism spectrum disorder (ASD). Said article got reported by several media outlets soon after its publication (e.g., Gallagher, 2014), and is likely to fuel further discussion on the link between autism and pollution.

It can be argued that both these studies, in character with most medical science investigations, were conducted with a view to inform policies. (As a matter of fact, Raz et al. (2014) concludes that "air pollution is a modifiable risk factor for autism, and reduced exposure during pregnancy could lead to lower incidence of ASD and reduce the substantial, increasing economic burden of ASD on families and on society".) CDC's investigation proved to be notably impactful in that regard.

Nonetheless, although associative studies are adequate for exploratory purposes and can be relied upon for the instatement of reasonable precautionary measures (e.g., efforts *should* be made to mitigate risk of Fusarium keratitis), they are generally inappropriate for policy making (recollection of ReNu with MoistureLoc®, *if* not a causative agent of Fusarium keratitis, is arguably unfair to Bausch & Lomb and its shareholders). In our view, policy makers should, instead, strive to rely on causal findings.

When analyzing case-noncase data for purposes of causal inference, we believe that an effort should be made to (a) work under a formal causal framework specifically tailored to retrospective designs, (b) state, within the confines of this framework, all assumptions made in reaching conclusions, and (c) of great importance, design the study, pre-analysis, without access to any outcome data. While conceptually straightforward for cohort studies, outcome-free design is complex for the case-noncase design. Any one-party, one-step design methodology without access to outcome data is infeasible: by design, sampling of units is conducted as a function of realized potential outcomes, which in turn induces dependence between treatment assignments and realized potential outcomes in the case-noncase sample (see Section 1.2.5). Consequently, this invalidates the naive implementation of matched sampling methods to case-noncase data. In Section 2.1, we propose a two-party, three-step methodology that circumvents this problem. To our knowledge, such methodology, along with a formal Rubin Causal Model-based framework for studying causal effects from case-noncase (or more generally, retrospective) data, has not been proposed to date.

1.2 A causal framework for case-noncase studies

In this section we introduce a formal framework for studying causal effects from case-noncase data, which extends the well-established Rubin Causal Model to retrospective studies.

1.2.1 The case-noncase study as a cohort study with missing data

We frame the case-noncase study as a (hypothetical) prospective cohort study with missing data (see Figure 1.1 below). Specifically, the data missing from this (hypothetical) cohort consists of covariate, treatment assignment, and potential outcome data for the non-sampled noncases — the non-sampled units, — and non-realized potential outcome data for all cases and those sampled noncases — the sampled units. We believe that this missing data formulation has the benefit of conceptually tying retrospective studies to their underlying prospective cohort studies, for which there exists an array of well-accepted methods for objective causal effect estimation (e.g., see Rubin, 2006; Imbens and Rubin, 2015). In Sections 1.2.2 through 1.2.8, we expound our missing data framework, under which our methodology, introduced in Chapter 2, operates.

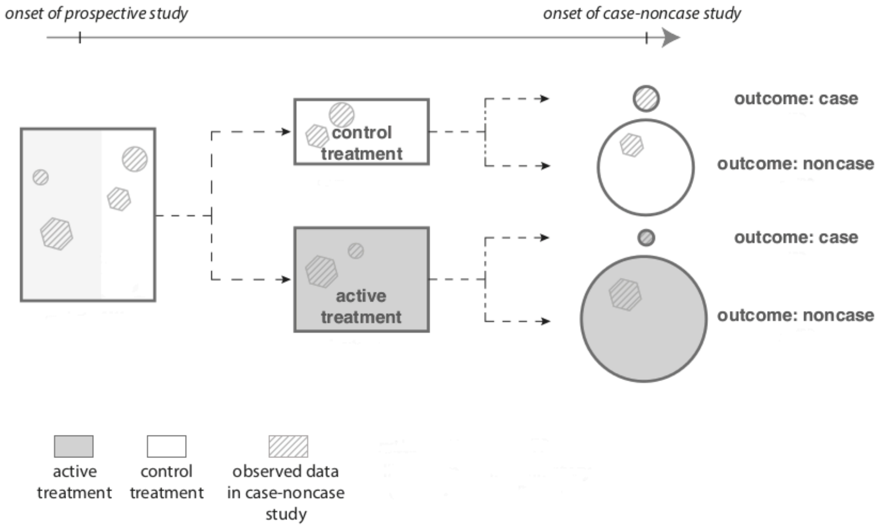

Figure 1.1: *The case-noncase study as a cohort study with missing data. Adapted from Greenberg et al. (2004).*

To be clear, the idea of the case-noncase study as a missing data problem is not new: this view was advocated by Wacholder in a 1996 Epidemiology paper (see Wacholder, 1996). Our work, however, departs from Wacholder's (and others'), in that (a) we focus on laying a formal framework for drawing causal inferences from case-noncase study data which involves, but is not limited to, missing data theory, (b) we put forward a methodology that enables objective estimation of causal effects from case-noncase data, and (c) our approach relies on multiple imputation, and is inherently Bayesian.

In effect, the inferential approach taken in Section 2.2 of this thesis is that of "calibrated Bayes" (Little, 2006). That is, while our method for inference is — in part — Bayesian, its properties are evaluated under the frequentist paradigm. In addition, our inferential framework is phenomenological, in that it focuses on observable values. Rubin (1978b), arguing in favor of this approach, notes that "[t]here do not exist parameters except under hypothetical models; there do, however, exist actual observed values and values that would have been observed. Focusing on the estimation of parameters is often not what the applied person wants to do since a hypothetical model is simply a structure that guides him to do sensible things with observed values." Lastly, in the spirit of the Rubin Causal Model, throughout this thesis we separate well-defined, observable objects of inference (e.g., the finite-population average causal effect), from the process by which the investigator learns about said objects of inference (e.g., the case-noncase study as a prospective observational study design with missing data), from the assumptions and statistical methods employed to estimate said quantities of interest (e.g., unconfoundedness, Bayesian multiple imputation). This approach contrasts with the commonly used techniques in the epidemiological literature (see Section 1.1.2) which generally, from the onset, define estimands as parameters embedded in some posited statistical model, without placing particular emphasis on question definition nor framework setup (nor pre-analysis design).

1.2.2 Population cohort and sample cohort

We define two notions — *population cohort* and *sample cohort* — to distinguish between two central sets of units, to which we allude throughout this thesis. We define *population cohort*[3] as the prospective cohort study population from which case-noncase data is sampled retrospectively, if such population exists (e.g., under a case-cohort design). Otherwise, we define the population cohort as the *hypothetical*

[3] The term *source population* has instead been previously used in various epidemiology papers. We believe that the term *population cohort* has the advantage of directly alluding to the sampling nature of the case-noncase problem.

prospective cohort study population from which case-noncase data is assumed to have arisen. Analogously, we define *sample cohort* as the set of units retrospectively sampled from the population cohort — this is the case-noncase sample.

The population cohort is the finite population of inference. That is to say, it is generally the population of interest to the investigator. For that reason, we do not attempt to infer anything about units outside this population in Chapter 2 of this thesis, which focuses on finite-population inference. Consequently, the process by which the population cohort is selected, e.g., by taking a simple random sample from some larger (super-)population or by virtue of availability of census data or hospital records, is immaterial to the finite-population analysis, but for the instance in which a super-population model is assumed for inferential purposes, as will hold true in Section 2.2.

1.2.3 Population cohort: complete (observable) data

The framework introduced here builds upon the well-established Rubin Causal Model for prospective cohort studies. The concept of unit, and notions of treatment, covariate, and potential outcomes — all of which are observable quantities — all hold under our framework, and apply to units in the population cohort. We supplement these ideas with the concept of *potential sampling indicators*, which we define below.

We begin by introducing notation, which readers acquainted with the RCM should find familiar. Let the population cohort consist of N units[4], indexed by $i = 1, \cdots, N$. Associated with each unit is a $1 \times k$ vector of covariates, $\boldsymbol{x}_i$. Let $\boldsymbol{X}$ denote the

[4] Henceforth, unless specified otherwise, “unit” is understood to mean “population cohort unit”.

$N \times k$ matrix consisting of all units' covariates. Further, let $\boldsymbol{W}$ denote the vector of assignments. The components W_i of $\boldsymbol{W}$ indicate exposure to active treatment when equal to 1, and zero otherwise. Under SUTVA (Holland, 1986), $Y_i(0)$ and $Y_i(1)$ denote the potential outcome values of unit i under control treatment and active treatment, respectively. We let $Y_i(w) = 1$ if unit i is a case when $W_i = w$, and $Y_i(w) = 0$ otherwise. Collectively, variables $\boldsymbol{Y}(1)$ and $\boldsymbol{Y}(0)$ constitute the $N-$vectors of potential outcomes, under active treatment and control treatment, respectively.

Drawing an analogy to potential outcomes, we introduce the idea of *potential sampling indicators.* The two potential sampling indicators associated with each unit are the indicators for inclusion of the unit in the sample cohort under each of the two treatment regimes. That is, we define unit i's *potential sampling indicators* as the sampling indicator that would be realized under control treatment and the sampling indicator that would be realized under active treatment. We denote the two variables by $S_i(0)$ and $S_i(1)$, respectively, and let $S_i(w) = 1$ when unit i is included in the cohort sample if assigned to treatment w, and 0 otherwise. For example, $S_3(0) = 1$ and $S_3(1) = 0$ signifies that unit 3 would be sampled under control treatment, but not under active treatment. Note that, because only one treatment can be applied to any given unit at any given time, only the potential sampling indicator corresponding to said applied treatment is realized; the other, namely, the potential sampling indicator corresponding to the alternative treatment, is missing. Also note that our definition tacitly assumes that the potential sampling indicator for any unit does not vary with treatments assigned to other units. We call this assumption SUTVA-S, analogously and in reference to SUTVA (Holland, 1986):

Assumption 1 (Stable Unit Treatment Value Assumption for Sampling (SUTVA-S))**.** The potential sampling indicators for any unit do not vary with the treatments assigned to other units.

More generally, for each unit i in the population cohort and treatment assignment vector $\boldsymbol{W}$, we let variable $S_i(\boldsymbol{W})$ denote the potential sampling indicator of unit i under treatment assignment vector $\boldsymbol{W}$. We then let $S_i(\boldsymbol{W}) = 1$ if unit i is sampled under assignment configuration $\boldsymbol{W}$, and $S_i(\boldsymbol{W}) = 0$ otherwise. This generalized notation would be used, for example, in a study in which sampling under allocation of 10% of units to active treatment, say, is of interest. SUTVA-S, however, does hold under the case-noncase setup considered throughout this thesis. We therefore let $\boldsymbol{S}(1)$ and $\boldsymbol{S}(0)$ denote the column potential sampling indicator vectors under active treatment and control treatment, respectively.

Ultimately, observable population cohort data consists of (a) the vectors of potential sampling indicators $\boldsymbol{S}(0)$ and $\boldsymbol{S}(1)$, (b) the vector of assignments $\boldsymbol{W}$, (c) the vectors of potential outcomes $\boldsymbol{Y}(0)$ and $\boldsymbol{Y}(1)$, and lastly (d) the matrix of covariates $\boldsymbol{X}$. We represent these variables jointly by the complete data matrix $\tilde{\boldsymbol{Y}}^{\text{compl}}$:

$$\tilde{\boldsymbol{Y}}^{\text{compl}} = (\boldsymbol{S}(0), \boldsymbol{S}(1), \boldsymbol{W}, \boldsymbol{Y}(0), \boldsymbol{Y}(1), \boldsymbol{X}) \tag{1.1}$$

1.2.4 Causal estimands

A causal effect is defined as the comparison of potential outcomes under active treatment and control treatment. A finite-population causal *estimand* is any function of the triplet $(\boldsymbol{Y}(0), \boldsymbol{Y}(1), \boldsymbol{X})$ that satisfies the definition of causal effect.

Let $\bar{Y}(0) = \frac{1}{N}\sum_i Y(0)$ and $\bar{Y}(1) = \frac{1}{N}\sum_i Y(1)$. In Table 1.1 below, we define three causal estimands of public health interest: the risk difference, the relative risk, and the odds ratio.

Table 1.1: *Three primary causal estimands for dichotomous outcomes. Adapted from Chretien (2010).*

Epidemiological term	**Finite population estimand**	**Super-population estimand**
Risk Difference (RD)	$\tau_{FP} = \bar{Y}(1) - \bar{Y}(0)$	$\tau_{SP} = E_{SP}[\tau_{FP}]$ $= \Pr(Y(1) = 1) - \Pr(Y(0) = 1)$
Relative Risk (RR)	$rr_{FP} = \frac{\bar{Y}(1)}{\bar{Y}(0)}$	$rr_{SP} = E_{SP}[rr_{FP}] \approx \frac{\Pr(Y(1)=1)}{\Pr(Y(0)=1)}$
Odds Ratio (OR)	$\omega_{FP} = \frac{\frac{\bar{Y}(1)}{1-\bar{Y}(1)}}{\frac{\bar{Y}(0)}{1-\bar{Y}(0)}}$	$\omega_{SP} = E_{SP}[\omega_{FP}] \approx \frac{\Pr(Y(1)=1)}{\Pr(Y(1)=0)} \Big/ \frac{\Pr(Y(0)=1)}{\Pr(Y(0)=0)}$

Estimands can be extended so as to incorporate covariate information. For example, an analyst may be interested in the risk difference for all males in the population cohort,

$$\tau_{males} = \frac{1}{N} \sum_{i\text{: unit } i \text{ is a male}} \Big(Y_i(1) - Y_i(0)\Big),$$

or in the effect of a drug for those units who were exposed to it:

$$\tau_{\text{treated}} = \frac{1}{N} \sum_{i:W_i=1} \Big(Y_i(1) - Y_i(0)\Big).$$

1.2.5 Two mechanisms create missing data, or not

Two mechanisms create missing data, or not; one via assignment of each unit to one of the two possible treatment regimes, the other via sampling or non-sampling of units into the sample cohort. We call these two mechanisms the assignment mechanism and the realized sampling mechanism, respectively.

Notation

To reflect the existence of two stages of missing data generation, we introduce double-superscript notation inspired by that in Rubin (1987). Under this notation, and when applicable[5], the first superscript indicates whether a given unit's measurement or variable (e.g., potential outcome) is realized or missing as a result of the assignment mechanism, and the second superscript indicates whether the unit in question is included in, or excluded from, the sample cohort.

We define this notation in Table 1.2 below, where Y denotes a dummy variable associated with unit i in the population cohort.

[5] See below for examples of non-applicability.

Table 1.2: *Notation.*
Y denotes a dummy variable associated with unit i in the population cohort.

Notation	**Definition**
$Y_i^{\text{r, }\cdot}$	realized (irrespective of inclusion or not in the cohort sample)
$Y_i^{\text{mis, }\cdot}$	missing (irrespective of inclusion or not in the cohort sample)
$Y_i^{\text{r, inc}}$	realized and included in the sample cohort
$Y_i^{\text{r, exc}}$	realized and excluded from the sample cohort
$Y_i^{\text{mis, inc}}$	missing and included in the sample cohort
$Y_i^{\text{mis, exc}}$	missing and excluded from the sample cohort
$Y_i^{\cdot\text{, inc}}$	always-realized (by default) and included in the sample cohort
$Y_i^{\cdot\text{, exc}}$	always-realized (by default) and excluded from the sample cohort

Accordingly, we let $Y_i^{\text{r, }\cdot}$ and $Y_i^{\text{mis, }\cdot}$ denote, respectively, the realized and missing potential outcome of each unit i in the population cohort. By definition:

$$Y_i^{\text{r, }\cdot} = W_i Y_i(1) + (1 - W_i) Y_i(0) \tag{1.2}$$

$$Y_i^{\text{mis, }\cdot} = (1 - W_i) Y_i(1) + W_i Y_i(0) \tag{1.3}$$

Likewise, associated with each unit i are one realized, and one missing, potential sampling indicator, which we denote by $S_i^{\text{r, }\cdot}$ and $S_i^{\text{mis, }\cdot}$, respectively. By definition, and under SUTVA-S:

$$S_i^{\text{r, }\cdot} = W_i S_i(1) + (1 - W_i) S_i(0) \tag{1.4}$$

$$S_i^{\text{mis, }\cdot} = (1 - W_i) S_i(1) + W_i S_i(0) \tag{1.5}$$

For example, $S_7^{\mathrm{r},\,\cdot} = 1$ would denote inclusion of unit 7 in the cohort sample under treatment received by unit 7. Note that, by default, the treatment assignment and covariate variables are "always-realized" in that they do not have missing potential counterparts the way potential outcomes and potential sampling indicators do. As such, we let $\boldsymbol{x}_i^{\cdot,\,\mathrm{inc}}$ and $\boldsymbol{x}_i^{\cdot,\,\mathrm{exc}}$ denote, respectively, the (always-realized and) included, and (always-realized and) excluded covariate vectors for unit i. Namely,

$$\boldsymbol{x}_i^{\cdot,\,\mathrm{inc}} = \boldsymbol{x}_i \quad \text{if} \quad S_i^{\mathrm{r},\,\cdot} = 1 \tag{1.6}$$

and $\boldsymbol{x}_i^{\cdot,\,\mathrm{exc}}$ is missing data when $S_i^{\mathrm{r},\,\cdot} = 0$. Similarly, we let $W_i^{\cdot,\,\mathrm{inc}}$ and $W_i^{\cdot,\,\mathrm{exc}}$ denote, respectively, the (always-realized and) included, and (always-realized and) excluded assignment for unit i. Hence,

$$W_i^{\cdot,\,\mathrm{inc}} = W_i \quad \text{if} \quad S_i^{\mathrm{r},\,\cdot} = 1 \tag{1.7}$$

and $W_i^{\cdot,\,\mathrm{exc}}$ is missing data when $S_i^{\mathrm{r},\,\cdot} = 0$. Finally, we let $Y_i^{\mathrm{r},\,\mathrm{inc}}$ and $Y_i^{\mathrm{mis},\,\mathrm{inc}}$ denote, respectively, unit i's realized and included, and missing and included potential outcome. That is,

$$Y_i^{\mathrm{r},\,\mathrm{inc}} = Y_i^{\mathrm{r},\,\cdot} \quad \text{if} \quad S_i^{\mathrm{r},\,\cdot} = 1 \tag{1.8}$$

$$Y_i^{\mathrm{mis},\,\mathrm{inc}} = Y_i^{\mathrm{mis},\,\cdot} \quad \text{if} \quad S_i^{\mathrm{r},\,\cdot} = 1 \tag{1.9}$$

Variables $Y_i^{\mathrm{r},\,\mathrm{exc}}$ and $Y_i^{\mathrm{mis},\,\mathrm{exc}}$ are excluded, and therefore constitute missing data. Note that these variables are non-identifiable without unit i's treatment assignment information.

Unit-level notation introduced in this section readily extends to vectors, and matrices when applicable.

The assignment mechanism

The first missing data mechanism, previously discussed in Section 1.1.3, is the *assignment mechanism.* One of the three central components of the RCM, the assignment mechanism is the process that governs which population cohort units are exposed to the active treatment, and which are exposed to the control treatment. Formally, it is defined (see Rubin, 1975; Imbens and Rubin, 2015) as a "function that assigns probabilities to all 2^N possible $N-$vectors of assignments $\boldsymbol{W}$, given the $N-$vectors of potential outcomes $\boldsymbol{Y}(0)$ and $\boldsymbol{Y}(1)$, and given the $N \times K$ matrix of covariates $\boldsymbol{X}$:

Definition 1 (Assignment Mechanism)**.** The assignment mechanism is a row-exchangeable function $\Pr(\boldsymbol{W}|\boldsymbol{W}, \boldsymbol{Y}(0), \boldsymbol{Y}(1))$, taking on values in $[0, 1]$, satisfying

$$\sum_{\boldsymbol{W} \in \{0,1\}^N} \Pr(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1)) = 1, \tag{1.10}$$

for all $\boldsymbol{X}, \boldsymbol{Y}(0)$, and $\boldsymbol{Y}(1)$."

We refer the reader to Imbens and Rubin (2015) for additional discussion and examples.

The potential sampling mechanism

As its name suggests, the potential sampling mechanism is the process that governs sampling or non-sampling of population cohort units under both treatment regimes. Formally, we define it as a function that assigns probabilities to all $2^N \times 2^N$ possible pairs of $N-$vectors of potential sampling indicators, given the $N-$vectors of potential outcomes $\boldsymbol{Y}(0)$ and $\boldsymbol{Y}(1)$, the $N \times K$ matrix of covariates $\boldsymbol{X}$, and given the $N-$vector of treatment assignments $\boldsymbol{W}$:

Definition 2 (Potential Sampling Mechanism)**.** Given a population cohort of N units, the potential sampling mechanism is a row-exchangeable function $\Pr(\boldsymbol{S}(0), \boldsymbol{S}(1) | \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1))$, taking on values in $[0, 1]$, satisfying

$$\sum_{\boldsymbol{S}(0) \in \{0,1\}^N,\ \boldsymbol{S}(1) \in \{0,1\}^N} \Pr(\boldsymbol{S}(0), \boldsymbol{S}(1) | \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1)) = 1, \tag{1.11}$$

for all $\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}(0)$, and $\boldsymbol{Y}(1)$.

Note that the potential sampling mechanism applies to both potential sampling indicator vectors despite it being impossible, by the fundamental problem of causal inference, to observe both indicators simultaneously for any given unit. This attribute makes it a hypothetical construct: under our case-noncase setup, for instance, whether any given unit would have been sampled under the treatment alternative to the one actually received[6] has no direct bearing on the (non-)inclusion of said unit in the cohort sample. Notwithstanding, we introduce the potential sampling mechanism as above for three main reasons.

[6] We assume treatment compliance and thereby use "assigned" and "received" interchangeably.

The first is pedagogical and expository. The potential sampling mechanism allows investigators to contemplate the occurrence (and to quantify the probability) of unit sampling under both treatment regimes: "Had she been exposed to the alternative treatment, would unit 5 have been selected into the sample cohort? If so, with what probability?". The mechanism, as such, elucidates the complex relationship between the sample cohort and the population cohort. What's more, we believe that, much in the same way that potential outcomes, in conjunction with the assignment mechanism, make explicit the very nature — and challenges — of causal inference (e.g., it being a missing data problem, where missingness is governed by the assignment mechanism), potential sampling indicators, in conjunction with the potential sampling mechanism, reveal the conceptually subtle nature of the causal inference problem for retrospective designs.

The second reason is that of generality. The potential sampling mechanism defined as above encompasses a large class of retrospective designs. Though fictional, a study design under which sampling of any given unit under any given treatment depends on that unit's missing potential outcome can be formally defined under our setup.

Lastly, the potential sampling mechanism serves as a natural building block for the *realized sampling mechanism*, which we define next.

The realized sampling mechanism

The realized sampling mechanism is the process that governs which population cohort units are selected into the sample cohort, and which are not. Formally, we define it as a function that assigns probabilities to all 2^N possible realized sampling vectors $\boldsymbol{S}^{r,\,\cdot}$, given the $N-$vectors of potential outcomes $\boldsymbol{Y}(0)$ and $\boldsymbol{Y}(1)$, the $N \times K$ matrix of covariates $\boldsymbol{X}$, and given the $N-$vector of treatment assignments $\boldsymbol{W}$:

Definition 3 (Realized Sampling Mechanism)**.** Given a population cohort of N units, the realized sampling mechanism is a row-exchangeable function $\Pr(\boldsymbol{S}^{\mathrm{r},\,\cdot}|\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1))$, taking on values in $[0, 1]$, satisfying

$$\sum_{\boldsymbol{S}^{\mathrm{r},\,\cdot} \in \{0,1\}^N} \Pr(\boldsymbol{S}^{\mathrm{r},\,\cdot}|\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1)) = 1, \tag{1.12}$$

for all $\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}(0)$, and $\boldsymbol{Y}(1)$.

Properties of the realized sampling mechanism

Imbens and Rubin (2015) discuss "three general properties that assignment mechanisms may satisfy": individualisticness, probabilisticness, and unconfoundedness. Below, we define analogues of said properties for the realized sampling mechanism. We then assess whether these properties hold under the case-noncase setup outlined in Section 1.1.1; namely, under sampling of all realized population cohort cases, and simple random sampling with known probability $\pi > 0$ of realized population cohort noncases.

(i) Individualisticness

Definition 4 (Individualistic realized sampling mechanism)**.** A realized sampling mechanism is individualistic if the realized sampling probability of any particular population cohort unit is only a function of that unit's assignment, covariate, and potential outcome values.

Proposition 1 (Individualisticness of realized sampling mechanism)**.** *The realized sampling mechanism is individualistic.*

Proof. The proposition follows immediately from the fact that, independently for $i = 1, \cdots, N$,

$$S_i^{\mathrm{r},\,\cdot} \overset{d}{=} Y_i^{\mathrm{r},\,\cdot} + B(1 - Y_i^{\mathrm{r},\,\cdot}) \tag{1.13}$$

where $B \sim \mathrm{Bern}(\pi)$ is independent of $Y_i^{\mathrm{r},\cdot}$.

■

(ii) Probabilisticness

Definition 5 (Probabilistic realized sampling mechanism)**.** A realized sampling mechanism is probabilistic if the realized sampling probability is strictly between zero and one for every unit in the population cohort. That is, every unit has the possibility of being selected into the sample cohort and the possibility of not being selected into the sample cohort.

Proposition 2 (Probabilisticness of realized sampling mechanism)**.** *The realized sampling mechanism is not probabilistic.*

Proof. The proposition follows from noting that, for those units for which $Y_i^{\mathrm{r},\,\cdot} = 1$,

$$\Pr(S_i^{\mathrm{r},\,\cdot} = 1 | W_i, X_i, Y_i(0), Y_i(1)) = 1. \tag{1.14}$$

■

(iii) Unconfoundedness

Definition 6 (Unconfounded realized sampling mechanism)**.** A realized sampling mechanism is unconfounded if it does not depend on the potential outcomes:

$$\Pr(\boldsymbol{S}^{\mathrm{r},\ \cdot}|\boldsymbol{W},\boldsymbol{X},\boldsymbol{Y}(0),\boldsymbol{Y}(1)) = \Pr(\boldsymbol{S}^{\mathrm{r},\ \cdot}|\boldsymbol{W},\boldsymbol{X},\boldsymbol{Y}'(0),\boldsymbol{Y}'(1)), \tag{1.15}$$

for all $\boldsymbol{W},\boldsymbol{X},\boldsymbol{Y}(0),\boldsymbol{Y}(1),\boldsymbol{Y}'(0)$, and $\boldsymbol{Y}'(1)$.

Proposition 3 (Confoundedness and ignorability of realized sampling mechanism)**.** *The realized sampling mechanism is confounded, but ignorable (Little and Rubin, 2002) in that it can be written as a function of* $\boldsymbol{W}$, $\boldsymbol{X}$ *and* $\boldsymbol{Y}^{r,\ \cdot}$ *only, without dependence on* $\boldsymbol{Y}^{mis,\cdot}$*.*

Proof. See Proposition 1 proof. ■

Corollary 1 (Ignorability of realized sampling mechanism for Bayesian inference)**.** *Under the conditions of Proposition 3, and assuming a priori independence between the parameters of the distributions of* $S_i^{r,\ \cdot}$ *and* $Y_i^{r,\ \cdot}$*, the realized sampling mechanism is ignorable for Bayesian inference, as defined in Little and Rubin (2002).*

We refer the reader to Little and Rubin (2002) for an in-depth discussion of ignorability. A key implication and benefit of an ignorable missing data mechanism is that it allows Bayesian inference to be based on the observed data likelihood only: the missing data mechanism, as the name suggests, can be ignored. Conveniently, the realized sampling mechanism *is* ignorable. However, it is also confounded.

Consider equation 1.13 above. By definition, for those units i in the sample cohort, $S_i^{\mathrm{r},\,\cdot} = 1$. By (1.13), therefore,

$$\begin{aligned} 1 &= S_i^{\mathrm{r},\,\cdot} && (1.16)\\ &\overset{d}{=} Y_i^{\mathrm{r},\,\cdot} + B(1 - Y_i^{\mathrm{r},\,\cdot}) && (1.17)\\ &= W_iY_i(1) + (1 - W_i)Y_i(0) + B(1 - [W_iY_i(1) + (1 - W_i)Y_i(0)]). && (1.18) \end{aligned}$$

This illustrates that confoundedness of the realized sampling mechanism induces distributional dependence between treatment assignments and (realized) potential outcomes in the sample cohort. In other words, the "sample cohort assignment mechanism" (i.e., the population cohort assignment mechanism, restricted to sample cohort units) is generally *not* unconfounded:

$$\Pr(\boldsymbol{W}^*|\boldsymbol{X}^*, \boldsymbol{Y}^*(0), \boldsymbol{Y}^*(1)) \neq \Pr(\boldsymbol{W}^*|\boldsymbol{X}^*) \qquad (1.19)$$

where for given variable $\boldsymbol{Z}$, $\boldsymbol{Z}^* := \boldsymbol{Z}|_{i:S_i^{\mathrm{r},\,\cdot}=1}$. (Potential outcomes cannot be dropped from the right-hand side of the equation, trivially because of the definition of starred variables). This invalidates the use of matched sampling methods on sample cohort data for purposes of pretreatment bias reduction in the estimation of *population cohort* causal effects.

1.2.6 Sample cohort: observed data

Ultimately, data that is observed by the analyst consists of (always-)realized variables for those units in the sample cohort. Specifically, observed data consists of (a) the vector of realized sampling indicators, $\boldsymbol{S}^{\text{r}, \ \cdot}$, (b) the vector of realized and included potential outcomes, $\boldsymbol{Y}^{\text{r, inc}}$, (c) the vector of included assignments , $\boldsymbol{W}^{\cdot, \text{ inc}}$ and lastly (d) the matrix of included covariates, $\boldsymbol{X}^{\cdot, \text{ inc}}$. We represent these variables jointly by the observed data matrix $\tilde{\boldsymbol{Y}}^{\text{obs}}$:

$$\tilde{\boldsymbol{Y}}^{\text{obs}} = (\boldsymbol{S}^{\text{r}, \ \cdot}, \boldsymbol{Y}^{\text{r, inc}}, \boldsymbol{W}^{\cdot, \text{ inc}}, \boldsymbol{X}^{\cdot, \text{ inc}}) \tag{1.20}$$

Henceforth, we shall refer to this set of observed data as *case-noncase data*, *realized sample cohort data*, or simply as *sample cohort data*.

Conversely, missing data consists of (a) the vector of missing and included potential outcomes $\boldsymbol{Y}^{\text{mis, inc}}$, (b) the vector of missing and excluded potential outcomes $\boldsymbol{Y}^{\text{mis, exc}}$, (c) the vector of excluded assignments, $\boldsymbol{W}^{\cdot, \text{ exc}}$ and (d) the matrix of excluded covariates, $\boldsymbol{X}^{\cdot, \text{ exc}}$. We represent these variables jointly by the missing data matrix $\tilde{\boldsymbol{Y}}^{\text{mis}}$:

$$\tilde{\boldsymbol{Y}}^{\text{mis}} = (\boldsymbol{Y}^{\text{mis, inc}}, \boldsymbol{Y}^{\text{r, exc}}, \boldsymbol{W}^{\cdot, \text{exc}}, \boldsymbol{X}^{\cdot, \text{exc}}) \tag{1.21}$$

Note that we omit $\boldsymbol{S}^{\text{mis}, \ \cdot}$ which, although important for conceptual understanding, plays no role in either missing data mechanism. Also note that under the assumption of sampling of all cases, $\boldsymbol{Y}^{\text{r, exc}}$ is known to be $\mathbf{0}$.

1.2.7 Sample cohort data generation, summarized

Sample cohort data generation can be compactly summarized in the form of Table 1.3 below. In this toy example, the population cohort consists of 24 units. Assumed is sampling of all cases, and simple random sampling of noncases with sampling probability $\pi = 0.5$. (Note that this sampling rate would generally be significantly lower in real-life studies.) As such, $Y_i(w) = 1$ implies $S_i(w) = 1$ for $w \in \{0, 1\}$, and all $i \in 1, \cdots, 24$. Bolded vectors refer to variables introduced throughout this section. Columns 2 to 5, in dark gray, represent observed data. Columns 7-12, in light gray, represent complete data. (In particular, columns 10-12 are what Rubin commonly refers to as "the science"; i.e. data which, if observed, would allow the analyst to directly calculate the causal estimand of her choosing.) Lastly, column 6, which consists of the vector $\mathbf{Y}^{\mathrm{r},\ \cdot}$, represents partially observed data. When read from right to left, the table illustrates sample cohort data generation, starting from the population cohort.

1.2.8 Additional terminology

We introduce additional terminology for purposes of expository clarity in Section 2.1 (see, in particular, Section 2.1.2 discussion), and clarity in Chapter 3. We define the *complete population cohort dataset* as the dataset consisting of population cohort covariates, potential outcomes, and treatment assignments. We define the *realized population cohort dataset* as the dataset consisting of population cohort covariates, realized potential outcomes, and treatment assignments.

Table 1.3: *Sample cohort data generation: a toy example with $N = 24$ population cohort units. All cases are sampled with probability 1; noncases are sampled via simple random sampling with probability $\pi = 0.5$. As such, $Y_i(w) = 1$ implies $S_i(w) = 1$ for $w \in \{0, 1\}$, and all $i \in 1, \cdots, 24$.*

Unit	$\boldsymbol{Y}^{\text{r, inc}}$	$\boldsymbol{W}^{\cdot\text{, inc}}$	$\boldsymbol{X}^{\cdot\text{, inc}}$	$\boldsymbol{S}^{\text{r, }\cdot}$	$\boldsymbol{Y}^{\text{r, }\cdot}$	$\boldsymbol{S}(0)$	$\boldsymbol{S}(1)$	$\boldsymbol{W}$	$\boldsymbol{Y}(0)$	$\boldsymbol{Y}(1)$	$\boldsymbol{X}$
1	1	0	$\boldsymbol{x}_1$	1	1	1	1	0	1	0	$\boldsymbol{x}_1$
2	1	0	$\boldsymbol{x}_2$	1	1	1	1	0	1	1	$\boldsymbol{x}_2$
3	1	0	$\boldsymbol{x}_3$	1	1	1	0	0	1	0	$\boldsymbol{x}_3$
4	1	0	$\boldsymbol{x}_4$	1	1	1	1	0	1	1	$\boldsymbol{x}_4$
5	1	1	$\boldsymbol{x}_5$	1	1	1	1	1	1	1	$\boldsymbol{x}_5$
6	1	1	$\boldsymbol{x}_6$	1	1	1	1	1	1	1	$\boldsymbol{x}_6$
7	1	1	$\boldsymbol{x}_7$	1	1	1	1	1	1	1	$\boldsymbol{x}_7$
8	1	1	$\boldsymbol{x}_8$	1	1	1	1	1	1	1	$\boldsymbol{x}_8$
9	0	0	$\boldsymbol{x}_9$	1	0	1	1	0	0	0	$\boldsymbol{x}_9$
10	0	0	$\boldsymbol{x}_{10}$	1	0	1	0	0	0	0	$\boldsymbol{x}_{10}$
11	0	0	$\boldsymbol{x}_{11}$	1	0	1	1	0	0	0	$\boldsymbol{x}_{11}$
12	0	0	$\boldsymbol{x}_{12}$	1	0	1	0	0	0	0	$\boldsymbol{x}_{12}$
13				0	0	0	1	0	0	0	$\boldsymbol{x}_{13}$
14				0	0	0	0	0	0	0	$\boldsymbol{x}_{14}$
15				0	0	0	1	0	0	0	$\boldsymbol{x}_{15}$
16				0	0	0	0	0	0	0	$\boldsymbol{x}_{16}$
17	0	1	$\boldsymbol{x}_{17}$	1	0	0	1	1	0	0	$\boldsymbol{x}_{17}$
18	0	1	$\boldsymbol{x}_{18}$	1	0	1	1	1	0	0	$\boldsymbol{x}_{18}$
19	0	1	$\boldsymbol{x}_{19}$	1	0	0	1	1	0	0	$\boldsymbol{x}_{19}$
20	0	1	$\boldsymbol{x}_{20}$	1	0	1	1	1	0	0	$\boldsymbol{x}_{20}$
21				0	0	1	0	1	0	0	$\boldsymbol{x}_{21}$
22				0	0	0	0	1	0	0	$\boldsymbol{x}_{22}$
23				0	0	1	0	1	1	0	$\boldsymbol{x}_{23}$
24				0	0	1	0	1	1	0	$\boldsymbol{x}_{24}$

1.2.9 On retrospective matching

As an immediate application of the framework thus introduced, we discuss the tangential topic of retrospective matching in case-noncase studies, from the causal inference perspective. Following Holland and Rubin (1988), let ‘prospective matching’ be understood as matching in the sense of Rubin (2005b); namely, the matching of control treatment units to active treatment units, on $\boldsymbol{X}$. Let retrospective matching[7] be understood as the “pairing of one or several noncases to each case, on the basis of their similarity with respect to selected variables” (Schlesselman, 1982).

Retrospective matching is used for noncase sampling in the ‘matched case-noncase’ study design. It is a popular practice in applied epidemiological research: 46.4% of case-noncase studies published in 1994 were of the matched type (Gefeller et al., 1998). In spite of its popularity, retrospective matching has long been the subject of controversy. For instance, according to a literature review by Rose and van der Laan (2009b), many early texts described the method as a way to reduce ‘confounding’, which Schlesselman defines as “the effect of an extraneous variable that wholly or partially accounts for the apparent effect of the study exposure, or that masks an underlying true association” (see, e.g., Schlesselman, 1982; Miettinen, 1970; Breslow and Powers, 1978; Breslow and Day, 1980). The more recent articles, however, have argued otherwise (see, e.g. Costanza, 1995; Rothman et al., 2008; Rose and van der Laan, 2009b).

Of note, Schlesselman (1982) writes that “the primary objective of matching is

[7] The term was coined by Holland and Rubin (1988), and is typically simply referred to as ‘matching’ in the epidemiological literature.

the elimination of biased comparisons between *cases* and [*noncases*]". Our position is that matching, on the contrary, should be used to ensure unbiased comparison between *control treatment* units and *active treatment* units. Prospective matching achieves the latter purpose by reconstructing the blocked randomized design within an observational dataset (Rubin, 2005b). That is, it creates balance in covariate distributions between the two treatment subgroups. Retrospective matching, as we now show, generally does not. Let us consider the example case-noncase datasets represented in Tables 1.4 and 1.5 below. Also, let us assume, for the sake of argument, unconfoundedness of the assignment mechanism in the presence of covariate *sex*.

By construction of dataset 1, exact retrospective matching is feasible — i.e, every case can be matched to a noncase based on *sex*. However, it is impossible for the investigator to learn about the treatment effect via prospective matching, because treatment assignment and *sex* are perfectly correlated. This shows that retrospective matching does not, generally, ensure comparability between control treatment and active treatement units.

Conversely, in example 2, exact prospective matching is possible — i.e., every control unit can be matched to an active unit based on *sex* — whereas retrospective matching is not — i.e., realized potential outcomes and *sex* are perfectly correlated. This suggests that retrospective matching can misguide the investigator into discarding data that is (highly) informative of the treatment effect.

Table 1.4: *Retrospective matching, example 1.*
'?' denotes a missing potential outcome.
'M' stands for male; 'F' for female.
Retrospective matching is possible.
Can't learn about treatment effect, even assuming unconfounded A.M.

Unit	$\boldsymbol{Y}(0)$	$\boldsymbol{Y}(1)$	$\boldsymbol{W}$	$\boldsymbol{Y}^{\text{r},\,\cdot}$	**sex**
1	1	?	0	1	M
2	1	?	0	1	M
3	0	?	0	0	M
4	0	?	0	0	M
5	?	1	1	1	F
6	?	1	1	1	F
7	?	0	1	0	F
8	?	0	1	0	F

Table 1.5: *Retrospective matching, example 2.*
'?' denotes a missing potential outcome.
'M' stands for male; 'F' for female.
Retrospective matching is impossible.
Can learn about treatment effect, assuming unconfounded A.M.

Unit	$\boldsymbol{Y}(0)$	$\boldsymbol{Y}(1)$	$\boldsymbol{W}$	$\boldsymbol{Y}^{\text{r},\,\cdot}$	**sex**
1	1	?	0	1	M
2	1	?	0	1	M
3	?	1	1	1	M
4	?	1	1	1	M
5	0	?	0	0	F
6	0	?	0	0	F
7	?	0	1	0	F
8	?	0	1	0	F

Another disadvantage of the use of retrospective matching is the potential for population cohort definition complication, or ill-definedness.

1.3 Discussion

To conclude, the causal inference framework introduced in this chapter extends the many benefits of the Rubin Causal Model to retrospective studies. Through formulation of the case-noncase study as a cohort study with missing data, our approach fills a conceptual gap between (observational) prospective cohort studies and retrospective studies. Conceptual coherence ensues: a case-noncase study is a partially observed cohort study, which itself is a broken stratified randomized experiment. The problem of causal inference for retrospective studies is therefore conceptually identical to that for cohort studies: the challenge is to reconstruct, to the extent possible, the broken randomized experiment.

In our view, our approach provides a deeper understanding of the case-noncase study design and related causal inference problem, than do traditional methodologies. For instance, the classical two-way table (Table 1.3) is typically used to summarize a case-noncase study, but generally provides no causal insight, as discussed in Holland and Rubin (1988). By contrast, our data generating mechanism table (Table 1.3) makes explicit (a) the relationship between the population cohort and sample cohort, and (b) the missing-data nature of the causal inference problem.

Moreover, as shown in Section 1.2.9, the potential outcomes perspective can shed new light on an age-old controversy.

Advantages of our framework are further discussed in Chapter 4.

Table 1.6: *Table of counts based on n_{+1} cases and n_{+0} noncases.*

	Noncases ($Y = 0$)	Cases ($Y = 1$)	Total
Control ($W = 0$)	n_{c0}	n_{c1}	n_{c+}
Treated ($W = 1$)	n_{t0}	n_{t1}	n_{t+}
Total	n_{+0}	n_{+1}	n_{++}

Chapter 2

PrepDA for objective causal inference

2.1 PrepDA: preprocessing, design and analysis of case-noncase study data for objective causal inference

2.1.1 Introducing PrepDA

As stated in Section 1.2.5, because of the realized sampling mechanism's confoundedness, and the thereby induced relationship between treatment assignments and (realized) potential outcomes in the sample cohort, it is invalid to use matched sampling methods (such as Mahalanobis metric matching) on sample cohort data for purposes of pretreatment bias reduction in the estimation of population cohort causal effects. As previously noted by Månsson et al. (2007), for instance, subclassification by propensity score estimates from all population cohort cases and a simple random sample of the noncases "should give consistent estimates of the true propensity score under the null hypothesis, but not otherwise". While one could analytically adjust sample cohort-based estimates of the population cohort propensity score by accounting for the implicit conditioning on $\{\boldsymbol{S}^{\mathrm{r},\ \cdot} = \mathbf{1}\}$, such approach would violate Rubin's principle of outcome-free design for causal effect estimation.

Instead, we propose a methodology that both circumvents the above confoundedness-related problem and adheres to Rubin's principles of objective design. The procedure, which we call by the acronym "PrepDA" for *Preprocessing Design Analysis*, stochastically recreates, via multiple imputation, a set of realized population cohort datasets from realized sample cohort data. Each such simulated dataset constitutes a prospective observational study, to which Rubin's outcome-free design and analysis procedures are then applied. This enables objective estimation of causal effects from case-noncase data via outcome-free matching or subclassification, post-data preprocessing. By preprocessing, we mean all contemplating, collecting, organizing, modeling, and imputation of data.

PrepDA relies on the existence of two parties, say 'A' and 'B', that operate independently from one another. For example, both parties could be research statisticians. In another example, party A would consist of a team of pharmaceutical biostatisticians and party B of independent statistical consultants.

Our methodology can be compactly summarized by the following three steps:

1. [***preprocessing***] Using sample cohort data, Party A multiply imputes, under

some statistical model, missing covariate, potential outcome, and treatment assignment data for the entire population cohort, thereby generating M simulated complete population cohort datasets. Party A then strips outcome data from each of the M imputed datasets. (Information regarding steps 2-4 is at this stage withheld from Party A. Party B is assumed *not* to be involved in this first step.)

Table 2.1 below depicts one such imputed dataset, starting hypothetically from Table 1.3 data.

Table 2.1: *Example of singly imputed complete population cohort dataset, obtained by implementing step 1 of PrepDA using observed data from Table 1.3.*
Imputed data appear in red, italicized.

Unit	$\boldsymbol{W}$	$\boldsymbol{Y}(0)$	$\boldsymbol{Y}(1)$	$\boldsymbol{X}$
1	0	1	*(0)*	$\boldsymbol{x}_1$
2	0	1	*(1)*	$\boldsymbol{x}_2$
3	0	1	*(0)*	$\boldsymbol{x}_3$
4	0	1	*(0)*	$\boldsymbol{x}_4$
5	1	*(1)*	1	$\boldsymbol{x}_5$
6	1	*(1)*	1	$\boldsymbol{x}_6$
7	1	*(1)*	1	$\boldsymbol{x}_7$
8	1	*(1)*	1	$\boldsymbol{x}_8$
9	0	0	*(0)*	$\boldsymbol{x}_9$
10	0	0	*(0)*	$\boldsymbol{x}_{10}$
11	0	0	*(0)*	$\boldsymbol{x}_{11}$
12	0	0	*(0)*	$\boldsymbol{x}_{12}$
13	*(1)*	*(0)*	*(0)*	$(\boldsymbol{x}_{13})$
14	*(1)*	*(1)*	*(0)*	$(\boldsymbol{x}_{14})$
15	*(0)*	*(0)*	*(0)*	$(\boldsymbol{x}_{15})$
16	*(1)*	*(0)*	*(0)*	$(\boldsymbol{x}_{16})$
17	1	*(0)*	0	$\boldsymbol{x}_{17}$
18	1	*(0)*	0	$\boldsymbol{x}_{18}$
19	1	*(1)*	0	$\boldsymbol{x}_{19}$
20	1	*(0)*	0	$\boldsymbol{x}_{20}$
21	*(1)*	*(0)*	*(0)*	$(\boldsymbol{x}_{21})$
22	*(1)*	*(0)*	*(0)*	$(\boldsymbol{x}_{22})$
23	*(1)*	*(1)*	*(0)*	$(\boldsymbol{x}_{23})$
24	*(0)*	*(0)*	*(0)*	$(\boldsymbol{x}_{24})$

Table 2.2: *Table 2.1 dataset, with outcomes suppressed.*

Unit	$\boldsymbol{W}$	$\boldsymbol{X}$
1	0	$\boldsymbol{x}_1$
2	0	$\boldsymbol{x}_2$
3	0	$\boldsymbol{x}_3$
4	0	$\boldsymbol{x}_4$
5	1	$\boldsymbol{x}_5$
6	1	$\boldsymbol{x}_6$
7	1	$\boldsymbol{x}_7$
8	1	$\boldsymbol{x}_8$
9	0	$\boldsymbol{x}_9$
10	0	$\boldsymbol{x}_{10}$
11	0	$\boldsymbol{x}_{11}$
12	0	$\boldsymbol{x}_{12}$
13	*(1)*	$(\boldsymbol{x}_{13})$
14	*(1)*	$(\boldsymbol{x}_{14})$
15	*(0)*	$(\boldsymbol{x}_{15})$
16	*(1)*	$(\boldsymbol{x}_{16})$
17	1	$\boldsymbol{x}_{17}$
18	1	$\boldsymbol{x}_{18}$
19	1	$\boldsymbol{x}_{19}$
20	1	$\boldsymbol{x}_{20}$
21	*(1)*	$(\boldsymbol{x}_{21})$
22	*(1)*	$(\boldsymbol{x}_{22})$
23	*(1)*	$(\boldsymbol{x}_{23})$
24	*(0)*	$(\boldsymbol{x}_{24})$

2. [***design***] Party A turns over the M outcome-free imputed datasets to Party B which, in turn, designs each one of the datasets. Design may involve, but is not limited to, data trimming, matching or subclassification, and covariate balance assessment (see, for example, Rubin, 2006, 2008; Imbens and Rubin, 2015).

In our running example, the outcome-free dataset from Table 2.1 above would be turned over to party B.

3. [***analysis***] Once Party B's design phase is finalized, Party A hands over to party B *realized* outcome data from each of the M imputed datasets from step 1. Party B then analyzes each of the M imputed and matched or subclassified realized population cohort datasets according to a strict pre-specified protocol, and combines the M (sets of) results using Rubin's rules for Multiple Imputation (Rubin, 1987).

Table 2.3 below depicts the data that would be analyzed by Party B in this third step of PrepDA.

Table 2.3: *Table 2.2 dataset, with singly imputed realized potential outcomes* $\boldsymbol{Y}^{r,\,\cdot}$ *from Table 2.1 appended. (Or, equivalently, singly imputed realized population cohort dataset.)*

Unit	$\boldsymbol{Y}^{\mathrm{r},\,\cdot}$	$\boldsymbol{W}$	$\boldsymbol{X}$
1	1	0	$\boldsymbol{x}_1$
2	1	0	$\boldsymbol{x}_2$
3	1	0	$\boldsymbol{x}_3$
4	1	0	$\boldsymbol{x}_4$
5	1	1	$\boldsymbol{x}_5$
6	1	1	$\boldsymbol{x}_6$
7	1	1	$\boldsymbol{x}_7$
8	1	1	$\boldsymbol{x}_8$
9	0	0	$\boldsymbol{x}_9$
10	0	0	$\boldsymbol{x}_{10}$
11	0	0	$\boldsymbol{x}_{11}$
12	0	0	$\boldsymbol{x}_{12}$
13	*(0)*	*(1)*	$(\boldsymbol{x}_{13})$
14	*(0)*	*(1)*	$(\boldsymbol{x}_{14})$
15	*(0)*	*(0)*	$(\boldsymbol{x}_{15})$
16	*(0)*	*(1)*	$(\boldsymbol{x}_{16})$
17	0	1	$\boldsymbol{x}_{17}$
18	0	1	$\boldsymbol{x}_{18}$
19	0	1	$\boldsymbol{x}_{19}$
20	0	1	$\boldsymbol{x}_{20}$
21	*(0)*	*(1)*	$(\boldsymbol{x}_{21})$
22	*(0)*	*(1)*	$(\boldsymbol{x}_{22})$
23	*(0)*	*(1)*	$(\boldsymbol{x}_{23})$
24	*(0)*	*(0)*	$(\boldsymbol{x}_{24})$

2.1.2 Discussion

A parallel with prospective cohort and randomized studies

Note that the preprocessing step yields a stochastically generated approximation of the case-noncase study's underlying prospective cohort study. This approach extends Rubin's (2007) proposal that prospective observational studies approximate randomized experiments as closely as possible.

The importance of a two-party procedure

By design, PrepDA relies on two independent parties, where the first is exclusively responsible for the preprocessing step, and the second for the design and analysis steps. In particular, Party A should neither be aware of (i.e., is blinded to) party B's design and analysis protocols, nor the study's causal objectives. To insure this, information regarding steps 2-4 is withheld from Party A at the time of step 1 implementation. The purpose of this separation is to minimize, to the extent possible, analyst bias in causal effect estimation. It prevents, in particular, any given analyst from preprocessing cohort data in a manner that (could) deliberately impact(s) the study's overall findings. An example of such intentional manipulation is the fit of various imputation models via trial and error, with knowledge of subsequent design schemes and analysis protocols. We illustrate this idea in Section 2.1.2 below.

The design and analysis, post-preprocessing, of ***realized*** versus ***complete*** imputed population cohort datasets

To be clear, in the third step of the procedure, party B analyzes *realized*, not *complete*, multiply imputed population cohort datasets. This serves two purposes. The first is to recreate the analysis that would have been performed on the prospective realized population cohort datasets, had that data been available to the investigator — see Section 2.1.2 comment above. The second is to prevent the methodology from being solely dependent on Party A's imputation model: party B, in fact, imputes missing (population cohort) potential outcome data in the analysis step of PrepDA.

Note that correct coverage of our procedure relies on having congeniality of the imputation and analysis models (see Meng, 1977). In practice, this should not be of major concern, unless models used by party A and party B have major inconsistencies. In addition, the use of matching or subclassification methods in the design phase of PrepDA, which limits reliance on model assumptions in the analysis step of PrepDA, mitigates such risk. Also note that Party B's analysis step may be redundant *if* the imputer's (Party A's) model accurately estimates missing potential outcome data. This, in turn, may result in the introduction of noise in the overall estimation of causal effects. We nonetheless believe that the added benefit of investigator objectivity, which PrepDA guarantees, outweighs the risk of possible addition of analytical bias and/or variance.

Towards a new notion of covariate balance

PrepDA requires a new notion of covariate balance between active treatment and control treatment groups. Because the causal effect is estimated by combining estimates obtained from M imputed datasets, there is a need to develop metrics that adequately summarize balance over multiple datasets. This topic is postponed to Chapter 3 of this thesis. An example of metric that comes immediately to mind is the average standardized distance between active treatment and control treatment covariate distributions, where the average is taken over the M imputed datasets.

A litigation example

Consider the scenario in which a pharmaceutical company is facing a class-action lawsuit alleging one of its drugs, say 'D', causes birth defects. Suppose, further, that evidence presented to the court includes a case-noncase dataset, which the plaintiffs' experts analyzed using logistic regression, and concluded a statistically significant association between intake of drug D and birth defects.

Now suppose that company's counsel engages two independent statistical consultants, 'A' and 'B', to analyze the dataset using PrepDA. Following consultant A's implementation of the preprocessing step, and upon consultant B's review of the M simulated outcome-free population cohort datasets, consultant B opines that there is not enough overlap in the distributions of key covariates between individuals who took drug D and those who did not take the drug, to conclude existence, or non-existence, of a causal effect of taking drug D on birth defects — the analysis, according to her, is inconclusive.

The above example illustrates a powerful application of PrepDA in the litigation setting. In contrast to plaintiffs' expert, analyst B could not have seen the litigation answer before making any assessments regarding adequacy of data and/or analysis methodology. Reliance on PrepDA, as such, protects her from accusations of deliberate data manipulation or subjective analysis, and thereby accords plausibility to her findings. We refer the reader to Kousser (1984) and Greiner (2008) for further discussion of the (mis)use of logistic regression in the litigation and academic settings. Another related reference is Robertson (2010), which discusses blind expertise in litigation.

Conclusion

To conclude, PrepDA shares the many advantages of Rubin's methodologies for causal inference (see Imbens and Rubin, 2015). Among other, PrepDA (a) allows for outcome-free matching or subclassification for pre-treatment bias reduction, post-preprocessing, (b) prevents researchers from running multiple (regression) models on observed data, and ultimately choosing the one result that is most in line with their research agenda, and (c) forces investigators to assess the validity of their causal findings by checking for overlap in covariate distributions between treatment groups, also post-preprocessing. We further discuss these advantages in Chapter 4.

2.2 Simulation study: estimation of τ_{FP} and ω_{FP} from case-noncase data, assuming presence of one covariate only

In this section we illustrate the application of PrepDA in a simple non-trivial setting that assumes the presence of one covariate only; i.e., $\boldsymbol{X}_{N \times k} = \boldsymbol{X}_{N \times 1}$. We investigate our methodology's operating characteristics via simulation under various conditions, and compare them to those of logistic and probit regression when applicable. The problem addressed throughout this section is that of causal inference with case-noncase data, where estimands of interest are the population cohort risk difference, $\tau_{FP} = \bar{Y}(1) - \bar{Y}(0)$, and the population cohort odds ratio, $\omega_{FP} = \frac{\frac{\bar{Y}(1)}{1-\bar{Y}(1)}}{\frac{\bar{Y}(0)}{1-\bar{Y}(0)}}$.

Naturally, we work under the framework introduced in Section 1.2.

2.2.1 A model for population cohort data generation

For each simulation condition (see Section 2.2.4), we generate complete population cohort data, $\tilde{\boldsymbol{Y}}^{\text{compl}}$, according to the following model. Let $\mu_0, \mu_1 \in \mathbb{R}$ and $p \in (0, 1)$. Marginally, the assignment of each unit i is modeled independently with a Bernoulli(p) distribution:

$$W_i | p \sim \text{Bern}(p). \tag{2.1}$$

For each i, independently and conditionally on treatment received W_i, we let covariate X_i follow a Normal distribution with mean μ_{W_i} and variance $\sigma^2_{W_i}$:

$$X_i | W_i, \mu_{W_i}, \sigma_{W_i} \sim \mathcal{N}(\mu_{W_i}, \sigma^2_{W_i}). \tag{2.2}$$

Specifications 2.1 and 2.2 induce the following assignment mechanism:

$$\Pr(W_i = w | X_i, \mu_0, \mu_1, \sigma_0, \sigma_1, p) = \frac{p^w (1-p)^{(1-w)} \cdot \phi(X_i; \mu_w, \sigma^2_w)}{(1-p) \cdot \phi(X_i; \mu_0, \sigma^2_0) + p \cdot \phi(X_i; \mu_1, \sigma^2_1)}, \tag{2.3}$$

for $w \in \{0, 1\}$. Next, let $\beta^{(j)}_0, \beta^{(j)}_X \in \mathbb{R}$ for $j \in \{0, 1\}$. For all i, and given covariates X_i, potential outcomes $Y_i(0)$ and $Y_i(1)$ are modeled independently according to generalized linear models:

$$\begin{aligned} \Pr(Y_i(0) = 1 | X_i, \beta^{(0)}_0, \beta^{(0)}_X) &= F(\beta^{(0)}_0 + \beta^{(0)}_X X_i), \text{ and} & (2.4) \\ \Pr(Y_i(1) = 1 | X_i, \beta^{(1)}_0, \beta^{(1)}_X) &= F(\beta^{(1)}_0 + \beta^{(1)}_X X_i), & (2.5) \end{aligned}$$

where $F(\cdot)$ is the c.d.f. of a specified distribution, and $F^{-1}(\cdot)$ is the link function. For purposes of this study, we consider the logit and probit links; see Section 2.2.4 for more details.

Finally, let $\pi \in (0, 1)$ fixed. We specify unit-level potential sampling probabilities

$$\Pr(S_i(0), S_i(1) | W_i, X_i, Y_i(0), Y_i(1), \pi)$$

independently for each unit i. We set $(S_i(0), S_i(1))$ to be conditionally independent of W_i and X_i (and $\beta_0^{(0)}, \beta_X^{(0)}, \beta_0^{(1)}$, and $\beta_X^{(1)}$) given $Y_i(0)$, $Y_i(1)$ and π:

$$\Pr(S_i(0), S_i(1) | W_i, X_i, Y_i(0), Y_i(1), \pi) = \Pr(S_i(0), S_i(1) | Y_i(0), Y_i(1), \pi) \tag{2.6}$$

Because sampling under control (active) treatment solely depends on unit i's potential outcome under control (active) treatment,

$$\Pr(S_i(0), S_i(1) | W_i, X_i, Y_i(0), Y_i(1), \pi) = \Pr(S_i(0) | Y_i(0), \pi) \cdot \Pr(S_i(1) | Y_i(1), \pi). \tag{2.7}$$

Together with the assumptions of sampling of all cases, and simple random sampling with known probability π of noncases, this yields:

$$\begin{aligned}\Pr(S_i(0) = s_{i0}, S_i(1) = s_{i1} | \cancel{W_i}, \cancel{X_i}, Y_i(0), Y_i(1), \pi) = & \left(\pi^{s_{i0}} (1-\pi)^{(1-s_{i0})}\right)^{1-Y_i(0)} \\ & \times \left(\pi^{s_{i1}} (1-\pi)^{(1-s_{i1})}\right)^{1-Y_i(1)}. \end{aligned} \tag{2.8}$$

We represent the set of parameters from the above model by $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = (p, \mu_0, \sigma_0, \mu_1, \sigma_1, \beta_0^{(0)}, \beta_X^{(0)}, \beta_0^{(1)}, \beta_X^{(1)})'. \tag{2.9}$$

2.2.2 Sample cohort data generation

For each simulation condition and generated complete cohort dataset $\tilde{\boldsymbol{Y}}^{\text{compl}}$, we obtain $\tilde{\boldsymbol{Y}}^{\text{obs}}$ via simple application of Section 1.2.5 definitions.

2.2.3 Two methods of analysis: PrepDA and regression adjustment

To each generated sample cohort dataset, we apply PrepDA and regression, as follows.

PrepDA

Given the computational intensity of our simulations, we implement an automated version of PrepDA. Practical demonstration of PrepDA's objectivity and its contrast to regression methods is postponed to Chapter 3.

1. [***preprocessing***] We generate $M = 100$ imputed population cohort datasets using Bayesian iterative simulation methods. We assume the generative model from Section 2.2.1 with probit link function $F^{-1}(\cdot) \equiv \Phi^{-1}(\cdot)$[1]. The following observed log-likelihood ensues:

[1] The probit link provides a convenient Gibbs sampler and closed-form analytical results.

$$
\begin{aligned}
\ell(\boldsymbol{\theta}|\tilde{\boldsymbol{Y}}^{\text{obs}}) = \sum_{i:S_i^{\text{r},\ \cdot}=1} \Bigg\{ & W_i^{\cdot,\ \text{inc}} Y_i^{\text{r, inc}} \log[\Phi(\beta_0^{(1)} + \beta_X^{(1)} X_i^{\cdot,\ \text{inc}})] \\
& + W_i^{\cdot,\ \text{inc}} (1 - Y_i^{\text{r, inc}}) \log[1 - \Phi(\beta_0^{(1)} + \beta_X^{(1)} X_i^{\cdot,\ \text{inc}})] \\
& + (1 - W_i^{\cdot,\ \text{inc}}) Y_i^{\text{r, inc}} \log[\Phi(\beta_0^{(0)} + \beta_X^{(0)} X_i^{\cdot,\ \text{inc}})] \\
& + (1 - W_i^{\cdot,\ \text{inc}})(1 - Y_i^{\text{r, inc}}) \log[1 - \Phi(\beta_0^{(0)} + \beta_X^{(0)} X_i^{\cdot,\ \text{inc}})] \\
& + (1 - W_i^{\cdot,\ \text{inc}}) \log[N(X_i^{\cdot,\ \text{inc}}; \mu_0, \sigma_0^2)] + W_i^{\cdot,\ \text{inc}} \log[N(X_i^{\cdot,\ \text{inc}}; \mu_1, \sigma_1^2)] \\
& + W_i^{\cdot,\ \text{inc}} \log[p] + (1 - W_i^{\cdot,\ \text{inc}}) \log[1 - p] \Bigg\} \\
& + (N - n_{\text{inc}}) \\
& \times \log\left[1 - (1-p) \cdot \Phi\left\{ \frac{\beta_0^{(0)} + \beta_X^{(0)} \mu_0}{\sqrt{1 + \sigma_0^2 {\beta_X^{(0)}}^2}} \right\} - \ p \cdot \Phi\left\{ \frac{\beta_0^{(1)} + \beta_X^{(1)} \mu_1}{\sqrt{1 + \sigma_1^2 {\beta_X^{(1)}}^2}} \right\} \right]
\end{aligned}
\tag{2.10}
$$

where $n_{\text{inc}} := \sum_i S_i^{\text{r},\ \cdot}$, $\boldsymbol{\theta}$ as in (2.9) and $\tilde{\boldsymbol{Y}}^{\text{obs}} = (\boldsymbol{S}^{\text{r},\ \cdot}, \boldsymbol{Y}^{\text{r, inc}}, \boldsymbol{W}^{\cdot,\ \text{inc}}, \boldsymbol{X}^{\cdot,\ \text{inc}})$.

Two key steps are repeated to distributionally impute missing data $\tilde{\boldsymbol{Y}}^{\text{mis}}$. The first consists of drawing a set of parameters $\boldsymbol{\theta}^*$ from the posterior distribution of the parameters given the observed data, $f(\boldsymbol{\theta}|\tilde{\boldsymbol{Y}}^{\text{obs}})$, using MCMC sampling via RStan software (Stan Development Team, 2014, see Appendix A.3.1 for further details). Given this draw for the parameters, we substitute the values $\boldsymbol{\theta}^*$ into the conditional distribution of $\tilde{\boldsymbol{Y}}^{\text{mis}}$ given $\tilde{\boldsymbol{Y}}^{\text{obs}}$ and $\boldsymbol{\theta}^*$, $f(\tilde{\boldsymbol{Y}}^{\text{mis}}|\tilde{\boldsymbol{Y}}^{\text{obs}}, \boldsymbol{\theta}^*)$, to impute a set of missing data $\tilde{\boldsymbol{Y}}^{\text{mis}}$. See Appendix A.3.2 for analytical and computational specifics involved in this second setup.

Next, we suppress both potential outcome vectors from each generated complete population cohort dataset.

2. [***design***] We trim each of the $M = 100$ outcome-free imputed population cohort datasets. That is, we discard those units in the active treatment arm whose covariate values do not overlap with the control treatment arm units' values, and those units in the control treatment arm whose covariate values do not overlap with the active treatment arm units' values. This prevents comparisons to be made between units that are too dissimilar in $\boldsymbol{X}$. We then partition the remaining units into 5 subclasses, $\mathcal{S} = 1, \cdots, 5$, using quantiles of $\boldsymbol{X}$.

3. [***analysis***] We analyze each of the M imputed realized population cohort datasets using two procedures, "Neyman Subclassification with Multiple Imputation and Trimming" (NSMIT) and "Haldane-Gart Subclassification with Multiple Imputation and Trimming" (HGSMIT). The first method estimates the population cohort risk difference, τ_{FP}, whereas the second one estimates the population cohort odds ratio, ω_{FP}. NSMIT and HGSMIT are summarized as follows:

   - Neyman Subclassification with Multiple Imputation and Trimming (NSMIT)

     For each trimmed imputed realized population cohort dataset (m), where $m = 1, \cdots, M$, we obtain a Neyman point estimate (Neyman, 1923) for the risk difference within each of the 5 subclasses $\mathcal{S}$:

$$\begin{aligned} \hat{\tau}_s^{(m)} &= \frac{1}{N_{ts}} \sum_{i:W_i=1,\ i \in \mathcal{S}} Y_i^{r,\cdot} - \frac{1}{N_{cs}} \sum_{i:W_i=0,\ i \in \mathcal{S}} Y_i^{r,\cdot} && (2.11) \\ &\equiv \hat{\tau}_{s,1}^{(m)} - \hat{\tau}_{s,0}^{(m)}, && (2.12) \end{aligned}$$

for $s \in \{1, \cdots, 5\}$, where $N_{ts} = \sum_{i=1,\ i \in \mathcal{S}}^{N} W_i$ and $N_{cs} = \sum_{i=1,\ i \in \mathcal{S}}^{N} (1 - W_i)$. We then obtain an estimate of the variance of $\hat{\tau}_s^{(m)}$:

$$\widehat{var(\hat{\tau}_s^{(m)})} = \frac{\hat{\tau}_{s,1}^{(m)}(1 - \hat{\tau}_{s,1}^{(m)})}{N_{ts}} + \frac{\hat{\tau}_{s,0}^{(m)}(1 - \hat{\tau}_{s,0}^{(m)})}{N_{cs}} \qquad (2.13)$$

where $N_s = \sum_{i \in \mathcal{S}} 1$. (Note that Neyman's method generates unbiased estimates of τ_{FP} and generally conservative intervals in large samples (Neyman, 1923; Imbens and Rubin, 2015).) Next, we compute the overall dataset (m)-specific Neyman point estimate, $\hat{\tau}^{(m)}$, by averaging across subclasses, weighting according to the number of units in each subclass, and the corresponding 95% large sample confidence interval.

We combine results from each of the M imputed datasets using Rubin's Rules for Multiple Imputation (Rubin, 1987) to get an overall estimate, $\bar{\tau}$, of risk difference:

$$\bar{\tau} = \frac{1}{M} \sum_m \hat{\tau}^{(m)} \qquad (2.14)$$

and an estimate, T, of its variance,

$$T \;=\; \Big(1+\frac{1}{M}\Big)B+\bar{U} \tag{2.15}$$

where $B=\frac{1}{M-1}\sum_m(\hat{\tau}^{(m)}-\bar{\tau})^2$ and $\bar{U}=\frac{1}{M}\sum_m \widehat{\mathrm{Var}(\hat{\tau}^{(m)})}$. The confidence interval for τ is obtained using:

$$(\bar{\tau}-\tau)/\sqrt{T}\sim t_\nu \;\text{ where }\; \nu=(M-1)\Big[1+\frac{\bar{\tau}}{(1+M^{-1})B}\Big]^2. \tag{2.16}$$

- <u>Haldane-Gart Subclassification with Multiple Imputation and Trimming (HGSMIT)</u>

  In a similar fashion to NSMIT, for each trimmed simulated realized population cohort dataset, we obtain an overall dataset (m)-specific point estimate for the log odds ratio by averaging subclass-specific point estimates. We then obtain the corresponding 95% interval. We use Haldane's extension (Haldane, 1955) of Woolf's method (Woolf, 1955) to estimate the log odds ratio within each subclass $\mathcal{S}$:

$$\log(\hat{\omega}_s^{(m)})=\log\Big[\frac{(n_s^{t1}+0.5)(n_s^{c0}+0.5)}{(n_s^{t0}+0.5)(n_s^{c1}+0.5)}\Big], \tag{2.17}$$

  for $s\in\{1,\cdots,5\}$, where $n_s^{t1}=\sum_{i=1,\ i\in\mathcal{S}}^N W_iY_i^{r,\cdot}$, $n_s^{c0}=\sum_{i=1,\ i\in\mathcal{S}}^N(1-W_i)(1-Y_i^{r,\cdot})$, $n_s^{t0}=\sum_{i=1,\ i\in\mathcal{S}}^N W_i(1-Y_i^{r,\cdot})$ and $n_s^{c1}=\sum_{i=1,\ i\in\mathcal{S}}^N(1-W_i)Y_i^{r,\cdot}$. (An advantage of Haldane's estimator over Woolf's is that the former exists for samples in which n_s^{t1}, n_s^{c0}, n_s^{t0} or n_s^{c1} has a null value.) We use Gart's method (Gart, 1966):

$$\log\hat{\omega} \pm 1.96\sqrt{1/(n_s^{t1}+0.5)+1/(n_s^{t0}+0.5)+1/(n_s^{c1}+0.5)+1/(n_s^{c0}+0.5)} \tag{2.18}$$

to compute 95% confidence intervals. (Note that Haldane's method produces an approximately unbiased estimate of ω_{FP}; Woolf's method generates generally conservative intervals in large samples (Haldane, 1955; Ding and Dasgupta, 2015).) We ultimately obtain an overall odds ratio estimate $\bar{\omega}$ and corresponding 95% confidence interval via application of Rubin's Rules for Multiple Imputation, followed by exponentiation.

(4.) [(***preprocessing results***)] In addition, we directly calculate the average risk differences and odds ratios from each of the M imputed complete population cohort datasets. This procedure yields M estimates of $\hat{\tau}_{FP}$, $\mathcal{T} = \{\hat{\tau}^{(1)}, \cdots, \hat{\tau}^{(M)}\}$, and M estimates of ω_{FP}, $\mathcal{O} = \{\hat{\omega}^{(1)}, \cdots .\hat{\omega}^{(M)}\}$. We obtain overall point estimates and credible intervals for τ_{FP} and ω_{FP} by taking the mean and (0.25, 0.975) quantiles of $\mathcal{M}$ and $\mathcal{O}$, respectively.

We henceforth refer to this method as "Multiple Imputation" (MI).

Logistic regression (LR) and probit regression (PR)

We fit logistic and probit regression models, where $\boldsymbol{Y}^{\text{r, inc}}$ is regressed on $\boldsymbol{W}^{\cdot,\text{ inc}}$ and $\boldsymbol{X}^{\cdot,\text{ inc}}$:

$$\text{logit}[\Pr(Y_i^{\text{r, inc}} = 1|X_i)] = \beta_0 + \beta_\omega W_i^{\cdot,\text{ inc}} + \beta_X X_i^{\cdot,\text{ inc}} \tag{2.19}$$

$$\Phi^{-1}[\Pr(Y_i^{\text{r, inc}} = 1|X_i)] = \beta_0 + \beta_\omega W_i^{\cdot,\text{ inc}} + \beta_X X_i^{\cdot,\text{ inc}} \tag{2.20}$$

In both cases, we exponentiate the estimated regression coefficient for $\boldsymbol{W}^{\cdot,\text{ inc}}$ to obtain an estimate $\hat{\omega}_{SP|W,X} = e^{\hat{\beta}_\omega}$ of ω_{FP}. We obtain a corresponding 95% confidence interval by exponentiating the endpoints of the conventional Normal-based confidence interval for $\hat{\beta}_\omega$.

Note that the logistic regression method provides estimates of the conditional, not marginal, and super, not finite, population odds ratio. Despite this, we implement it so as to investigate the method's performance in estimating the causal population cohort odds ratio, an estimand we believe is of greater interest in many, if not most, epidemiological studies.

Also note that probit regression is generally not used for purposes of (conditional) odds ratio inference. We implement it here to investigate the regression methods' sensitivity to link function $F^{-1}(\cdot)$ misspecification.

2.2.4 Simulation design

We investigate frequentist operating characteristics of PrepDA under various simulation conditions, and assess how logistic and probit regression perform in comparison for estimating ω_{FP}. Table 2.4 below specifies factors used in our study, which can be described as a 2^6 factorial design.

Table 2.4: *Simulation factors.*

Factor	**Levels of factor**
N	$\{1000\}$
π	$\{0.1\}$
p	$\{0.5\}$
$\beta_0^{(0)}$	{-3, -2}
$\beta_0^{(1)}$	{-3, -2}
$\beta_X^{(0)}$	{-1, -$\frac{1}{2}$}
$\beta_X^{(1)}$	{-1, -$\frac{1}{2}$}
σ_0^2	$\{2\}$
μ_0	$\{0\}$
σ_1^2	$\{1\}$
$B = \frac{\mu_t - \mu_c}{\sqrt{\frac{\sigma_1^2+\sigma_0^2}{2}}}$	$\{0, 1\}$
$F^{-(1)}(\cdot)$	{logit$(\cdot)$, $\Phi(\cdot)$}

In a typical case-cohort study, the two first factors would be known to the investigator, whereas the last ten would be unknown. Simulation parameters were selected so as to sensibly emulate real-life settings while taking into account the computationally intensive nature of the study. The latter, in particular, informed our choice of population cohort size, N. A sampling rate of noncases of $\pi = 0.1$, in conjunction with chosen levels of N and $\boldsymbol{\beta} := \left(\beta_0^{(0)}, \beta_X^{(0)}, \beta_0^{(1)}, \beta_X^{(1)}\right)$ ensured generation of sufficiently large sample cohorts for purposes of statistical inference. Factor F^{-1} investigates departure from the probit link, which is assumed in the preprocessing step of PrepDA. Lastly, as in Cochran and Rubin (1973), we parametrize the distance between treated and control group covariate means in terms of the standardized bias

$$B = \frac{\mu_t - \mu_c}{\sqrt{\frac{\sigma_1^2 + \sigma_0^2}{2}}} \tag{2.21}$$

so as to evaluate the factor's influence independently of the variance ratio $\frac{\sigma_0^2}{\sigma_1^2}$. A level of B=2 was initially considered, but yielded too little overlap in covariate distributions between treated and control sample cohort units, and was thus withdrawn from the study.

For every combination of factor levels above, we generate 100 complete population cohort datasets, as described in Section 2.2.1. From each generated complete population cohort dataset, we obtain a sample cohort dataset. For purposes of comparison with logistic and probit regression methods, data is regenerated if perfect separation occurs in the fit of either the logistic or probit model. We then perform six different analyses: MI and NSMIT to estimate the population cohort risk difference, τ_{FP}, and MI, HGSMIT, LR and PR to estimate the population cohort odds ratio, ω_{FP}. Each method is evaluated based on three criteria: mean coverage of the corresponding nominal 95% intervals[2], mean absolute percent bias, and mean interval width.

Our study can be compactly summarized in the form of a pseudo-algorithm, displayed below:

[2] To be clear, Bayesian posterior predictive intervals for MI and confidence intervals for NSMIT, HGSMIT, LR and PR.

Algorithm 1: Overview of simulation study

```
for each of the 64 simulation conditions do
    (a) generate 100 complete population cohort datasets;
    (b) generate 100 realized sample cohort datasets;
    repeat steps (a) & (b) if perfect separation occurs;
    for each realized sample cohort dataset do
        apply PrepDA as outlined in Section 2.2.3;
        run logistic and probit regressions as outlined in Section 2.2.3;
evaluate frequentist properties of each method by computing:
    mean coverage
    mean absolute percent bias
    mean interval length
```

2.2.5 Results and discussion

Table 2.5 below provides a summary of generated population cohort data.

Table 2.5: *Summary of generated population cohort data, across the* 64×100 *generated datasets.*

	Min	**1st quartile**	**Median**	**3rd quartile**	**Max**
$\frac{1}{N}\sum Y_i(0)$	< 0.001	0.030	0.056	0.103	0.206
$\frac{1}{N}\sum Y_i(1)$	< 0.001	0.030	0.056	0.104	0.198
τ_{FP}	-0.149	-0.030	0.000	0.030	0.143
ω_{FP}	0.000	0.4829	1.000	2.031	107.300

The number of units in our generated sample cohort datasets ranged from 80 to 282, with a median of 158.

Population cohort risk difference τ_{FP}

Mean coverage of nominal 95% intervals *(See Table 2.6).* MI has approximately 92-94% coverage in all conditions except when $B = 1$ and the probit link is used. In that case, it under-covers the true risk difference. Nominal 95% intervals generated by NSMIT have approximately 97%-98% coverage when data is generated using the logistic link. Otherwise, coverage varies both by levels of B and treatment effect. When $B = 0$, NSMIT yields approximately nominal coverage. The method under-covers when $B = 1$, with under-coverage being more significant in the presence of a treatment effect. We speculate this to be in part due to departures from the large-sample Normality assumptions that are required for the construction of Neymanian confidence intervals. In fact, under several simulation conditions, generated sample cohort datasets had as little as 100 units, 5 of which are cases. After subclassification, this entailed even smaller sample sizes. Also note that, under our simulation conditions, parameters $\boldsymbol{\theta}$ are not sampled from posited generative prior distributions. For this reason, small departures from nominal coverage are expected. Finally, note that for both methods, coverage is in general higher when the logistic link is used to generate data. This is as expected, given the slightly wider tails of the logistic data generative model.

Table 2.6: *Mean coverage of nominal 95% interval for τ.*
*No treatment effect is defined in terms of super-population parameters $\boldsymbol{\beta}$. That is, $\beta_0^{(0)} = \beta_0^{(1)}$ and $\beta_X^{(0)} = \beta_X^{(1)}$.

		No treatment effect*		**Treatment effect**	
Method	**B \ DGP link**	Probit	Logit	Probit	Logit
MI	0	.93	.94	.93	.92
	1	.83	.94	.84	.92
NSMIT	0	.96	.98	.94	.97
	1	.83	.98	.66	.97

Mean absolute percent bias ***(See Table 2.7).*** Under correct model specification, MI produces more bias when $B = 1$ than when $B = 0$. In all simulation settings, NSMIT produces less biased estimates than MI, as expected. In particular, under correct model specification and when $B = 1$, the difference between mean absolute percent bias produced by MI and NSMIT is most significant. Namely, there exists a 2.5 to 4-fold decrease in mean absolute percent bias when NSMIT is applied in lieu of MI. This demonstrates NSMIT's effectiveness in reducing pretreatment bias in the estimation of the risk difference estimand.

Table 2.7: *Mean absolute percent bias (τ estimand).*
*No treatment effect is defined in terms of super-population parameters $\boldsymbol{\beta}$. That is, $\beta_0^{(0)} = \beta_0^{(1)}$ and $\beta_X^{(0)} = \beta_X^{(1)}$.

		No treatment effect*		**Treatment effect**	
Method	**B \ DGP link**	Probit	Logit	Probit	Logit
MI	0	4.39	4.41	0.46	0.54
	1	12.17	4.41	2.15	0.54
NSMIT	0	3.96	4.37	0.42	0.53
	1	2.94	4.37	0.86	0.53

Mean nominal 95% interval width *(See Table 2.8).* Nominal 95% intervals are narrow, which is expected given the rarity of outcome. As expected, MI produces wider intervals when $B = 1$ than when $B = 0$: precision decreases when covariate distributions differ between the two treatment groups.

Table 2.8: *Mean width of nominal 95% interval for τ.*
*No treatment effect is defined in terms of super-population parameters $\boldsymbol{\beta}$. That is, $\beta_0^{(0)} = \beta_0^{(1)}$ and $\beta_X^{(0)} = \beta_X^{(1)}$.

		No treatment effect*		**Treatment effect**	
Method	**B \ DGP link**	Probit	Logit	Probit	Logit
MI	0	.06	.10	.06	.10
	1	.10	.10	.10	.10
NSMIT	0	.00	.00	.00	.00
	1	.00	.00	.00	.00

Population cohort odds ratio ω_{FP}

Mean coverage of nominal 95% intervals *(See Table 2.9).* 95% intervals generated by MI cover the odds ratio with approximately the same rate as for the risk difference estimand. 95% intervals generated by HGSMIT have approximately 98%-99% coverage under no treatment effect, approximately nominal coverage when the logistic link is used to generate data and a treatment effect exists, and coverage levels lower than 82% when the probit link is used to generate data and a treatment effect exists. LR produces approximately nominal coverage when the logistic link is used to generate data and a treatment effect exists. It slightly under-covers when the probit link is used and a treatment exists. Otherwise, under the null hypothesis of no treatment effect, LR typically over-covers the odds ratio. Not surprisingly, PR yields coverage of approximately 65% when a treatment effect exists (the method is not intended for odds ratio estimation), and 97% under the null hypothesis of no treatment effect. As is the case for the risk difference estimand, under logistic link data generation all methods generate intervals with coverage approximately equal to or greater than under probit link data generation. Also, coverage rates are generally higher under the null hypothesis of no treatment effect.

Table 2.9: *Mean coverage of nominal 95% interval for ω.*
*No treatment effect is defined in terms of super-population parameters $\boldsymbol{\beta}$. That is, $\beta_0^{(0)} = \beta_0^{(1)}$ and $\beta_X^{(0)} = \beta_X^{(1)}$.

		No treatment effect*		**Treatment effect**	
Method	**B \ DGP link**	Probit	Logit	Probit	Logit
MI	0	.93	.93	.92	.92
	1	.83	.93	.85	.92
HGSMIT	0	.99	.98	.82	.95
	1	.98	.98	.70	.95
LR	0	.98	.98	.94	.95
	1	.98	.98	.92	.95
PR	0	.98	.97	.66	.68
	1	.97	.97	.63	.68

Mean absolute percent bias *(See Table 2.10).* Similarly to the pattern observed for NSMIT in the estimation of risk difference, HGSMIT is most effective in reducing bias under correct model specification and when $B = 1$. In this setting, the estimates it generates are less biased than those generated by LR and PR. Under logistic link data generation, the HGSMIT estimator has slightly larger mean absolute percent bias than that of LR. This difference is negligible under the null hypothesis of no treatment effect (2% difference), and somewhat more prominent under non-null treatment effect (9% difference). We note that LR is more sensitive to model misspecification, in that HGSMIT's advantage over LR is greater under probit link data generation (with as much as a 6-fold reduction in bias) than is LR's advantage over HGSMIT under logistic data generation (with at most a 1.2-fold reduction in bias). Also, under correct link function specification, HGSMIT outperforms MI in all but the setting of a non-null treatment effect and when $B = 0$. While this exception requires further theoretical investigation, the former observation is expected, and is due to trimming and subclass-specific estimation of odds ratios in HGSMIT. In contrast, HGSMIT adds bias in the overall estimation of the odds ratio when the logistic link is used to generate data; that is, in situations in which MI seems to already be accurately estimating ω_{FP}. This was anticipated in Section 2.1.2. Lastly, note that while B significantly influences mean absolute percent bias under probit link data generation, it has no effect whatsoever when the logistic link is used.

Table 2.10: *Mean absolute percent bias (ω estimand).*
*No treatment effect is defined in terms of super-population parameters $\boldsymbol{\beta}$. That is, $\beta_0^{(0)} = \beta_0^{(1)}$ and $\beta_X^{(0)} = \beta_X^{(1)}$.

		No treatment effect*		**Treatment effect**	
Method	**B \ DGP link**	Probit	Logit	Probit	Logit
MI	0	0.72	0.23	0.67	0.25
	1	2.72	0.23	3.41	0.25
HGSMIT	0	0.30	0.25	1.11	0.34
	1	0.39	0.25	1.45	0.34
LR	0	0.59	0.23	$\gg 100$	0.29
	1	2.67	0.23	2.61	0.29
PR	0	0.31	0.14	0.86	0.32
	1	0.66	0.14	1.96	0.32

Mean nominal 95% interval width *(See Table 2.11).* When a probit link is used to generate data, both LR and PR yield notably wide confidence intervals.

This merits further theoretical investigation. With non-overlap between covariate distributions in the two treatment groups, HGSMIT produces narrower intervals than MI under correct model specification (i.e., probit link) and wider intervals otherwise. When data is generated via the logistic link, LR produces wider intervals than MI, but slightly narrower intervals than HGSMIT. The substantive effect of B on mean interval width is the same as it is on mean absolute percent bias. All four methods produce narrower intervals when the null hypothesis of no treatment effect holds.

Table 2.11: *Mean width of nominal 95% interval for ω.*
*No treatment effect is defined in terms of super-population parameters $\boldsymbol{\beta}$. That is, $\beta_0^{(0)} = \beta_0^{(1)}$ and $\beta_X^{(0)} = \beta_X^{(1)}$.

		No treatment effect*		**Treatment effect**	
Method	**B \ DGP link**	Probit	Logit	Probit	Logit
MI	0	3.73	1.15	7.95	1.52
	1	11.96	1.15	13.95	1.52
HGSMIT	0	3.76	1.72	5.92	2.36
	1	3.51	1.72	4.04	2.36
LR	0	13.91	1.48	72.86	2.08
	1	$\gg 100$	1.48	$\gg 100$	2.08
PR	0	$\gg 100$	0.82	$\gg 100$	0.93
	1	$\gg 100$	0.82	$\gg 100$	0.93

Summary

Our simulations show that NSMIT and HGSMIT generally yield reasonable coverage rates for nominal 95% intervals for τ and ω, although there are also indications of the methods' failure to perform satisfactory in more challenging situations (e.g., in the presence of covariate imbalance). Said methods' mean percent bias reduction properties are as expected: when there exists non-overlap in covariate distributions between the two treatment groups, NSMIT and HGSMIT are both effective in reducing bias, except when MI, in the preceding step 1 of PrepDA, produces accurate estimates of the risk difference and odds ratio, respectively. In these settings, NSMIT and HGSMIT are prone to introducing minor noise in the overall estimation of causal effects. Moreover, HGSMIT generally yields similar results to LR in our controlled simulation settings with the presence of one covariate only. HGSMIT (and MI), however, appears to be less sensitive to link function misspecification than LR and PR with regards to its bias reduction and mean interval width properties.

We expect PrepDA's analytical advantage over regression adjustment to better materialize in practical settings. For one, dangers of linear extrapolation can be significant with the latter method, especially in high dimensional observational studies. Step 2 of PrepDA should attenuate this problem. Second, regression-based data snooping — which PrepDA disallows — can substantially impact analysis results, as will be shown in Chapter 3.

Of note, Appendix A.2 outlines a Bayesian method for inferring causal effects from case-noncase data under the setup of framework 1.2, but with covariates suppressed. Our simulations show that this method has properties similar to those of Woolf's (1955) well-accpeted procedure for the estimation of ω_{SP}. In addition, our method produces credible intervals that achieve nominal 95% coverage.

On the application of PrepDA in practice

Because of the goal of our simulation study (to assess frequentist properties of PrepDA), its setting (presence of one covariate only), and the automated nature of the procedure, we did not in our algorithm check for overlap in covariate distributions between the two treatment groups. We advise that analysts perform this crucial step when implementing PrepDA in practice. Also note that the population to which causal findings apply can change if data trimming occurrs. This should also be taken into consideration in practical settings.

2.2.6 Extension to multivariate normal model

The model from 2.2.1 for population cohort data readily extends to multivariate normal $\boldsymbol{X}$, as follows. Let $p \in (0,1)$. Marginally, the assignment of each unit i is modeled independently with a Bernoulli(p) distribution:

$$W_i|p \sim \text{Bern}(p). \tag{2.22}$$

Let $\boldsymbol{\mu}_0 := (1, \mu_1^{(0)}, \cdots, \mu_k^{(0)})^T$, $\boldsymbol{\mu}_1 := (1, \mu_1^{(1)}, \cdots, \mu_k^{(1)})^T \in \mathbb{R}^k$, $\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1 \in \mathbb{S}_{++}^{k}$[3] and $p \in (0,1)$. Then, for all i, independently and conditionally on treatment received W_i, we let the $1 \times k$ vector of covariates $\boldsymbol{x}_i := (X_{i1}, \cdots, X_{ik})$ follow a multivariate Normal distribution with mean $\boldsymbol{\mu}_{W_i}$ and variance $\boldsymbol{\Sigma}_{W_i}$:

$$\boldsymbol{x}_i|W_i, \boldsymbol{\mu}_{W_i}, \boldsymbol{\Sigma}_{W_i} \sim \mathcal{MVN}(\boldsymbol{\mu}_{W_i}, \boldsymbol{\Sigma}_{W_i}) \tag{2.23}$$

[3] $\mathbb{S}_{++}^n = \{A \in \mathbb{R}^{n\times n} : A = A^t \text{ and } x^T Ax > 0 \text{ for all } x \in \mathbb{R}^n \text{ such that } x \neq 0\}$

Specifications 2.22 and 2.23 induce the following assignment mechanism:

$$\Pr(W_i = w|\boldsymbol{x}_i, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1, p) = \frac{p^w(1-p)^{(1-w)} \cdot \phi_k(\boldsymbol{x}_i; \boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)}{(1-p) \cdot \phi_k(\boldsymbol{x}_i; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + p \cdot \phi_k(\boldsymbol{x}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma})}, \tag{2.24}$$

for $w \in \{0,1\}$. Next, let $\boldsymbol{\beta}^{(j)} = (\beta_0^{(j)}, \cdots, \beta_k^{(j)})^T \in \mathbb{R}^{k+1}$ for $j \in \{0,1\}$, and $\widetilde{\boldsymbol{x}}_i := (1, X_{i1}, \cdots, X_{ik}) \in \mathbb{R}^{1\times(k+1)}$. For all i, and given the vector of covariates $\boldsymbol{x}_i$, potential outcomes $Y_i(0)$ and $Y_i(1)$ are modeled independently according to the following two probit models:

$$\Pr(Y_i(0) = 1|\boldsymbol{x}_i, \boldsymbol{\beta}^{(0)}) = \Phi(\widetilde{\boldsymbol{x}}_i \cdot \boldsymbol{\beta}^{(0)}) \tag{2.25}$$

$$\Pr(Y_i(1) = 1|\boldsymbol{x}_i, \boldsymbol{\beta}^{(1)}) = \Phi(\widetilde{\boldsymbol{x}}_i \cdot \boldsymbol{\beta}^{(1)}) \tag{2.26}$$

Finally, let the sampling mechanism as in equation (2.8).

Let the set of parameters from the above model be represented by $\boldsymbol{\theta}$, where

$$\boldsymbol{\theta} = (p, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}). \tag{2.27}$$

In addition, let $\widetilde{\boldsymbol{\mu}}_j := (1, \mu_1^{(j)}, \cdots, \mu_k^{(j)})^T \in \mathbb{R}^{k+1}$ for $j \in \{0,1\}$ and

$$\widetilde{\boldsymbol{\Sigma}}_W := \begin{vmatrix} 0 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \boldsymbol{\Sigma}_W & \\ 0 & & & \end{vmatrix}.$$

The observed log-likelihood is thus:

$$
\begin{aligned}
\ell(\boldsymbol{\theta}|\tilde{\boldsymbol{Y}}^{\text{obs}}) = \sum_{i:S_i^{\text{r},\,\cdot}=1} \Bigg\{ & W_i^{\cdot,\ \text{inc}} Y_i^{\text{r, inc}} \log[\Phi(\widetilde{\boldsymbol{x}}_i^{\cdot,\ \text{inc}} \boldsymbol{\beta}^{(1)})] \\
& + W_i^{\cdot,\ \text{inc}} (1 - Y_i^{\text{r, inc}}) \log[1 - \Phi(\widetilde{\boldsymbol{x}}_i^{\cdot,\ \text{inc}} \boldsymbol{\beta}^{(1)})] \\
& + (1 - W_i^{\cdot,\ \text{inc}}) Y_i^{\text{r, inc}} \log[\Phi(\widetilde{\boldsymbol{x}}_i^{\cdot,\ \text{inc}} \boldsymbol{\beta}^{(0)})] \\
& + (1 - W_i^{\cdot,\ \text{inc}})(1 - Y_i^{\text{r, inc}}) \log[1 - \Phi(\widetilde{\boldsymbol{x}}_i^{\cdot,\ \text{inc}} \boldsymbol{\beta}^{(0)})] \\
& + (1 - W_i^{\cdot,\ \text{inc}}) \log[\phi_k(\boldsymbol{x}_i^{\cdot,\ \text{inc}}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)] \\
& + W_i^{\cdot,\ \text{inc}} \log[\phi_k(\boldsymbol{x}_i^{\cdot,\ \text{inc}}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)] \\
& + W_i^{\cdot,\ \text{inc}} \log[p] + (1 - W_i^{\cdot,\ \text{inc}}) \log[1 - p] \Bigg\} \\
& + (N - n_{\text{inc}}) \\
& \times \log\left[1 - (1 - p) \cdot \Phi\left\{ \frac{\widetilde{\boldsymbol{\mu}}_0^T \boldsymbol{\beta}^{(0)}}{\sqrt{1 + \boldsymbol{\beta}^{(0)^T} \widetilde{\boldsymbol{\Sigma}}_0 \boldsymbol{\beta}^{(0)}}} \right\} - \ p \cdot \Phi\left\{ \frac{\widetilde{\boldsymbol{\mu}}_1^T \boldsymbol{\beta}^{(1)}}{\sqrt{1 + \boldsymbol{\beta}^{(1)^T} \widetilde{\boldsymbol{\Sigma}}_1 \boldsymbol{\beta}^{(1)}}} \right\} \right]
\end{aligned}
\tag{2.28}
$$

where $n_{\text{inc}} := \sum_i S_i^{\text{r},\ \cdot}$, $\boldsymbol{\theta}$ as in (2.27) and $\tilde{\boldsymbol{Y}}^{\text{obs}} = (\boldsymbol{S}^{\text{r},\ \cdot}, \boldsymbol{Y}^{\text{r, inc}}, \boldsymbol{W}^{\cdot,\ \text{inc}}, \boldsymbol{X}^{\cdot,\ \text{inc}})$.

A strategy, analogous to that presented in Section 2.2.3, for statistical inference under the above model, is given in Appendix A.3.

Chapter 3

PrepDA and logistic regression, contrasted: a reanalysis of data from Karkouti et al. (2006)

3.1 Objectives, background, and analysis outline

3.1.1 Objectives of this chapter

In a study analogous to LaLonde (1986) and Dehejia and Wahba (2002), we investigate whether PrepDA and logistic regression, when applied to case-noncase data, can generate estimates that are concordant with those from the causal analysis of population cohort data. The purpose of our work is to (a) illustrate the application of PrepDA in the context of a real-life example, (b) investigate consistency, or lack thereof, between results obtained via the application of PrepDA and logistic regression to *sample cohort* data, and those derived from the causal analysis of *population cohort* data, (c) introduce tools for covariate balance assessment across multiple imputed datasets, and (d) explore the potential for analyst bias with logistic regression, when said method is used to analyze case-noncase data. To this end, we focus on the re-analysis of a subset of data from a published article, Karkouti et al. (2006), which we detail below.

3.1.2 Example dataset: Karkouti et al. (2006)

Karkouti et al. (2006) concerns a prospective nonrandomized study of two drugs, aprotinin[1] and tranexamic acid[2], in patients who underwent cardiac surgery at the Toronto General Hospital from 1999 to 2004. Aprotinin and tranexamic acid are both used to prevent or treat excessive blood loss during complex surgery, such as cardiac surgery. Until 2006, aprotinin was generally considered to be superior to trenexamic acid, despite a lack of supporting clinical evidence (Linden, 2003; Karkouti et al., 2006). Karkouti et al. (2006) investigates the drugs' relative clinical utility and safety on a variety of outcomes, such as postoperative risk of blood product transfusion, stroke, infection, and mortality.

3.1.3 Construction of realized population cohort dataset

From data used in Karkouti et al. (2006), we construct a sub-dataset for purposes of our analysis. We select eight key covariates, and construct another four

[1] Trasylol, produced by Bayer AG, Toronto, Ontario, Canada.

[2] Cyclokapron, produced by Pharmacia & UpJohn Inc., Mississauga, Ontario, Canada.

by summing indicator variables related to the following clinical attributes: heart surgery history, clinical presentation, coronary artery disease risk, and coronary artery disease-associated illness. Table 3.1 details our resulting set of covariates. We consider as primary outcome *postoperative renal failure*, defined as new requirement for dialysis support[3], and selected for its rarity (approximate rate of 1.90%). Of note, we disregard patients with missing covariate or outcome data. The ensuing realized population cohort dataset comprises $N_{pop} = 7,416$ patients, of whom 407, or 5.49%, received aprotinin, and of whom $n_c = 141$ are cases (i.e., patients with postoperative renal failure).

[3] Dialysis is a process for removing waste and excess water from the blood, and is used primarily as an artificial replacement for lost kidney function in people with renal failure (Medicine Net Staff, 2014).

Table 3.1: *Description of covariates in constructed dataset.*
The first six covariates appear in the original dataset; that last four were constructed.
Endocarditis is an infection of the inner lining of the heart (Mayo Clinic Staff, 2014a).
†An *elective admission* is a surgery that is scheduled in advance, because it does not involve a medical emergency (Mosby, 2009).
‡*Platelets*, or thrombocytes, are colorless blood cells that play an important role in blood clotting. Platelets stop blood loss by clumping and forming plugs in blood vessel holes (Mayo Clinic Staff, 2014b).

Name	**Description**	**Levels (if applicable)** **or example**
type.surg	type of surgical procedure	isolated bypass, valve, other
act.endoc	indicator for active endocarditis*	none, remote, active, active abscess
pre.HB	preoperative hemoglobin (HB) concentration in dag/dL	
elective.surg	indicator for non-elective† admission	elective, admission
age	patient age	
sex	patient sex	female, male
area	patient body surface in m^2	
plt.count	platelet‡ count in $10^6/L$	
prev.surg	sum of indicators for previous heart surgeries	e.g., aortic valve surgery
clinical	sum of indicators for clinical presentation variables	e.g., most recent myocardial infarction
cad.risk	sum of indicators for coronary artery disease (CAD) risks	e.g., diabetes
asst.dis	sum of indicators for diseases associated with CAD	e.g., previous stroke

3.1.4 Generation of realized sample cohort dataset

From the aforementioned constructed dataset, we generate a synthetic case-noncase sample. Following the case-cohort study design, we sample all cases, and take a simple random sample of 10% of the noncases. The resulting realized sample cohort dataset consists of $N_{\text{sample}} = 868$ patients, of whom 76, or 8.76%, received aprotinin.

3.1.5 Outline of analysis strategy

Our strategy for data analysis is as follows. In Section 3.2, we perform the — benchmark — causal analysis of population cohort data. In Sections 3.3 and 3.4, we analyze sample cohort data using PrepDA and logistic regression, respectively. Estimands of interest are the super-population risk difference,

$$\tau_{SP} \quad = \quad \Pr(Y(1) = 1) - \Pr(Y(0) = 1), \tag{3.1}$$

and the super-population causal odds ratio,

$$\omega_{SP} \quad = \quad \frac{\Pr(Y(1) = 1)}{\Pr(Y(1) = 0)} \bigg/ \frac{\Pr(Y(0) = 1)}{\Pr(Y(0) = 0)}, \tag{3.2}$$

in Sections 3.2 and 3.3, and the super-population conditional associative odds ratio,

$$\omega_{SP|W,\boldsymbol{x}} \quad = \quad \frac{\Pr(Y(W) = 1|W = 1, \boldsymbol{x})}{\Pr(Y(W) = 0|W = 1, \boldsymbol{x})} \bigg/ \frac{\Pr(Y(W) = 1|W = 0, \boldsymbol{x})}{\Pr(Y(W) = 0|W = 0, \boldsymbol{x})}, \tag{3.3}$$

in Section 3.4.

Henceforth, we define “receipt of aprotinin” as active treatment, and “receipt of tranexamic acid" as control treatment. Considering the context of our study, SUTVA holds. In addition, we assume unconfoundedness of the assignment mechanism, despite having chosen for analysis only a subset of pretreatment variables considered by Karkouti et al.

3.2 Causal analysis of realized population cohort dataset

In this section, we estimate the causal effect of aprotinin versus tranexamic acid on postoperative renal failure, using the realized population cohort dataset. In doing so, we follow Rubin's guidelines for the design and analysis of observational studies (Rubin, 2007).

3.2.1 Design of observational data

Via logistic regression, we estimate propensity scores using all twelve covariates available in the dataset. We run a 1:1 greedy nearest neighbor matching algorithm (Rubin, 1973a) to select, out of a pool of 7009 control patients, 407 matches for treated patients, based on their proximity on propensity score distance. Figures 3.1-3.3 and Table 3.2 summarize the result of our algorithm. That is, matching yields an overall satisfactory balance in covariate distributions between active treatment and control treatment units. In particular, Figures 3.2 and 3.3 demonstrate an overlap in estimated propensity scores between the control and treated subpopulations, after matching. Table 3.2 and Figure 3.1 show a notable improvement in balance of covariate means for all 12 covariates, also after matching. In addition, *t*-tests (Table B.1, Appendix B.1) suggest non-significant differences in covariate means, post-matching, between the two treatment groups, at the 0.05 significance level for all covariates but *prev.surg*.

Efforts were not undertaken to prioritize a subset of covariates for matching, given the illustrative, as opposed to scientifically investigative, nature of our analysis.

Table 3.2: *Mean within each treatment group for each covariate, before and after matching.*
Note: *TA* stands for tranexamic acid.

	Initial		**After matching**	
	$\bar{X}_{aprtotinin}$	$\bar{X}_{TA}$	$\bar{X}_{aprotinin}$	$\bar{X}_{TA}$
type.surg	2.56	1.47	2.56	2.61
act.endoc	0.10	0.01	0.10	0.08
pre.HB	126.66	133.98	126.66	128.99
elective.surg	0.44	0.42	0.44	0.42
age	55.37	62.86	55.37	55.64
sex	1.37	1.26	1.37	1.38
area	0.19	0.03	0.19	0.19
plt.count	8.94	1.93	8.94	8.20
prev.surg	1.08	0.09	1.08	0.84
clinical	5.57	6.55	5.57	5.43
cad.risk	1.83	3.12	1.83	1.78
asst.dis	0.72	0.44	0.72	0.71

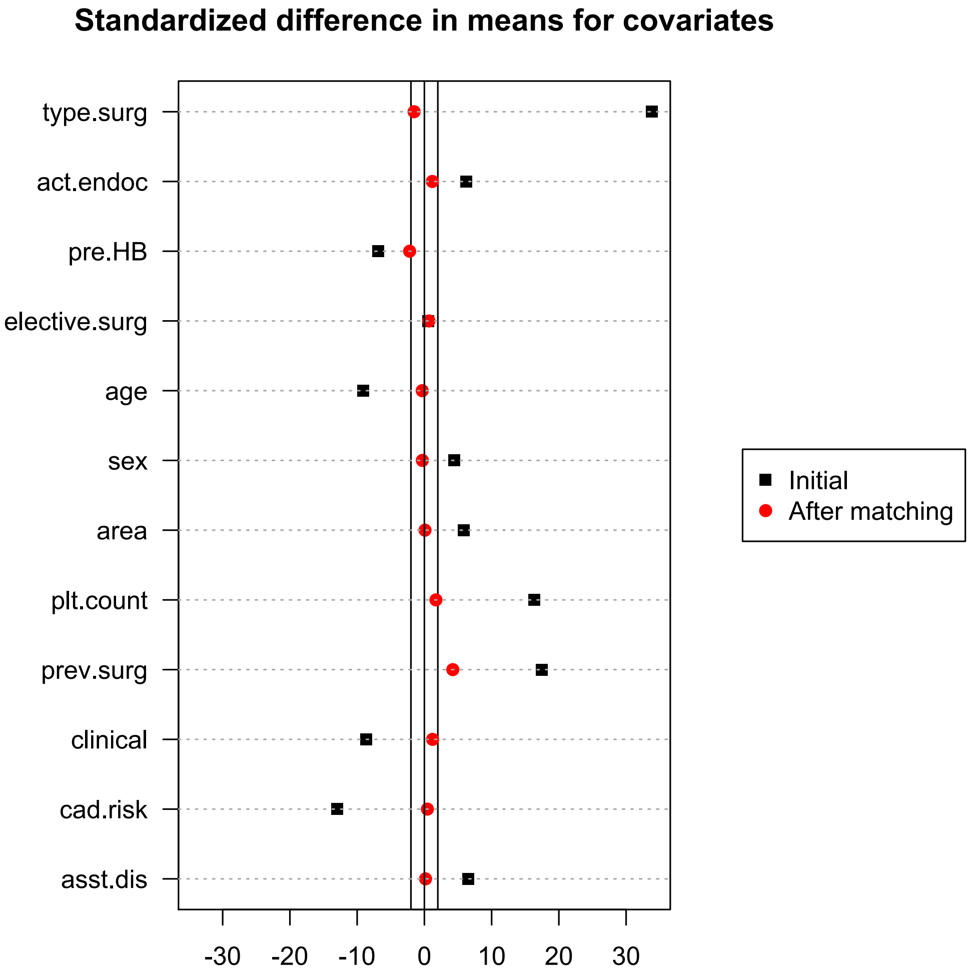

Figure 3.1: *Standardized difference in means, initial and after matching, for covariates.*
Note: to ensure fair before and after comparison, post-matching differences in means were standardized using the estimate of the variance of differences in means before matching. Displayed post-matching statistics are thus not t-statistics in the conventional sense.
Vertical lines appear at standardized differences in means of -2, 0 and 2, respectively.

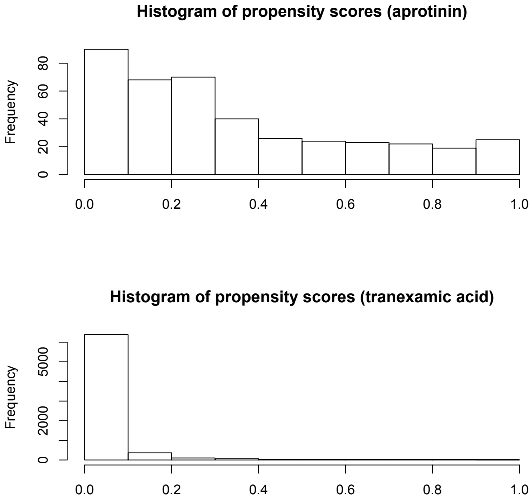

Figure 3.2: Histograms of estimated propensity scores, by treatment, before matching.

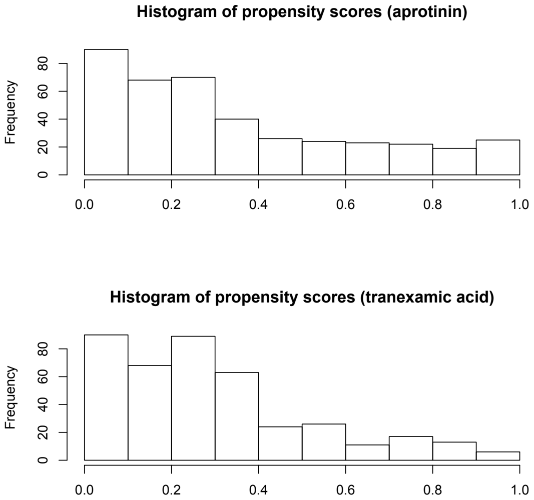

Figure 3.3: Histograms of estimated propensity scores, by treatment, after matching.

3.2.2 Analysis outline and results

We apply Neyman's method and simple linear regression on matched data to estimate τ_{SP}. We use Woolf's method[4] and logistic regression, also on matched data, to estimate ω_{SP}. In both regression models, *postoperative renal failure* is regressed on the indicator for receipt of aprotinin, and all twelve covariates. Regression is used to adjust for any residual imbalance in covariate distributions (e.g., on variable *prev.surg*) between the active treatment and control treatment subgroups (Rubin and Stuart, 2007). Rubin (1973b, 1979); Robins and Rotnitzky (1995); Heckman et al. (1997); Rubin and Thomas (2000) discuss the benefits of combining regression with matching. Our results are displayed in Table 3.3 below. Results from the application of the above methods on unmatched data are also included, for reference.

[4] Woolf's estimator and Haldane's (1955) approximately unbiased estimator produced approximately identical estimates of the odds ratio.

Table 3.3: *Causal analysis of realized population cohort dataset: results.*

		Unmatched group		**Matched group**	
estimand	**method**	**point estimate**	**95% CI**	**point estimate**	**95% CI**
risk difference	Neyman	0.071	(0.043, 0.098)	0.012	(-0.025, 0.050)
	Regression	0.017	(0.001, 0.032)	0.004	(-0.031, 0.038)
odds ratio	Woolf	6.13	(4.12, 9.10)	1.18	(0.71, 1.97)
	Regression	1.66	(0.95, 2.91)	1.34	(0.74, 2.44)

Our analysis concludes a non-significant causal effect of aprotinin versus tranexamic acid on postoperative renal failure at the 0.05 significance level, for both the risk difference and odds ratio estimands. Neyman and linear regression yield estimates of 1.2% and 0.4%, respectively, for the risk difference. Woolf and logistic regression estimate the odds ratio at 1.18 and 1.34, respectively. Given our use of well-established causal inference methods in the above analysis, we regard these findings as benchmarks.

3.3 Causal analysis of realized sample cohort dataset via PrepDA

In this section we estimate the causal effect of aprotinin versus tranexamic acid on postoperative renal failure, this time using realized *sample* cohort data. We implement the three steps of PrepDA (Section 2.1.1), barring the presence of two independent parties, to estimate τ_{SP} and ω_{SP}, as follows.

3.3.1 Step 1: preprocessing

Exploiting the fact that, within the confines of our study, population cohort data is known, we impute missing realized population cohort data,

$$\tilde{\boldsymbol{Y}}^{\text{mis}} = (\boldsymbol{Y}^{\text{r, exc}}, \boldsymbol{W}^{\cdot,\text{ exc}}, \boldsymbol{X}^{\cdot,\text{ exc}}),$$

by drawing from the empirical approximation of the *true* conditional distribution of missing data given the observed data, $\Pr(\tilde{\boldsymbol{Y}}^{\text{mis}}|\tilde{\boldsymbol{Y}}^{\text{obs}})$, where $\tilde{\boldsymbol{Y}}^{\text{obs}}$ is defined as in (1.20). We refer the reader to Appendix B.4 for further details. We thus generate a total of $M = 20$ imputed population cohort datasets[5].

This imputation strategy, although generally impracticable, ensures the implementation of PrepDA under correct imputation model specification. Furthermore, it circumvents the current limitations of our methodology; that is, to normally distributed covariates and probit link function specification. Also note that the method here circumvents the imputation of missing potential outcome data. In practice, we recommend that analysts posit an imputation model on the *complete* population cohort data matrix so as to adequately model all relevant data.

Figure 3.4 summarizes the result of our imputation procedure for covariate *pre.HB*. It depicts, for active treatment units, histograms of said covariate, in both the population cohort and sample cohort. It also depicts superimposed densities from all 20 imputed population cohort datasets. As expected, distributions of *pre.HB* differ in the population and sample cohort. This distortion is due to the confounded nature

[5] Our results were insensitive to additional imputations.

of the realized sampling mechanism. (See Figures B.1 and B.2 in Appendix B.2 for additional histograms.) What's more, we observe that the imputed *pre.HB* data is in agreement with "true", population cohort, *pre.HB* data. That is, our imputation procedure successfully re-creates a stochastic version of the realized population cohort, shown here for covariate *pre.HB*.

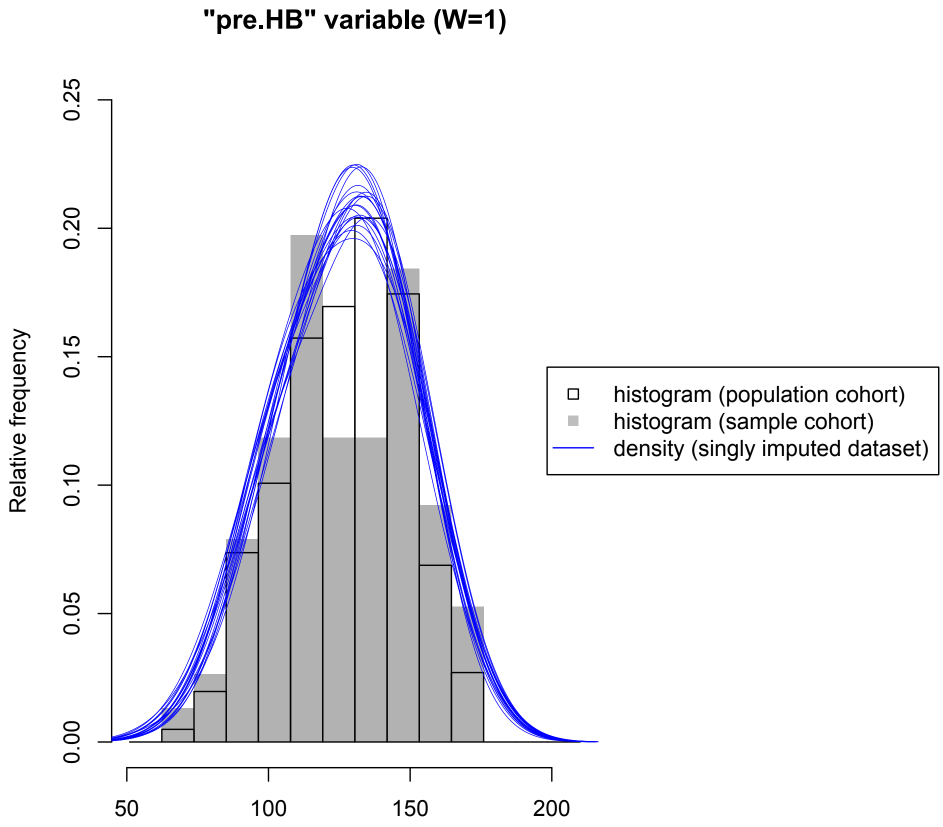

Figure 3.4: *Histogram/density of pre.HB variable in realized population cohort, realized sample cohort, and imputed realized population cohort.*

3.3.2 Design

We apply the matching algorithm from Section 3.2.1 to each imputed dataset. We then assess overall covariate balance. To this end, we construct two types of plots which, in our opinion, together effectively summarize balance of covariate means across imputed datasets. The first plot, pictured in Figure 3.5, is an extension of Figure 3.1. It displays standardized differences in means for all covariates in all imputed datasets, before and after matching. In our example, we note an improvement in average balance of covariate means, after matching, for all covariates. Further, with the exception of *prev.surg*, differences in standardized means are desirably concentrated around the null value. The second type of plot, in Figures 3.6 and 3.7, depicts (relative) changes in standardized differences in means, pre to post-matching, in all imputed datasets, for a given covariate. This plot conveys the share of datasets within which matching improved, or worsened, balance. In our example, balance improved significantly across all datasets for covariate *type.surg* in uniform fashion. Such is, however, the case in only 11 out of the 20 datasets for *elective.surg*. Nonetheless, *type.surg* is well-balanced post-matching.

Note that the two preceding plots can be amended to display alternative metrics, such as differences in covariate quantiles between the two treatment subgroups. Examples of additional balance metrics and visual balance diagnostics are:

- Metrics for assessment of overall balance:
  - Average, taken over M imputed datasets, mean Mahalanobis distance between covariate values of units in the active treatment and control treatment subgroups.
  - Proportion of covariates whose average balance improved upon matching.
- Metrics for assessment of balance for a given covariate:

  - Median difference, across imputed datasets, in covariate ranges between active treatment and control treatment units.
  - Proportion, out of those imputed datasets for which initial balance was unsatisfactory, of datasets with satisfactory balance after matching.

- Visual diagnostics:
  - Plot of superimposed densities, as in Figure 3.4, of estimated propensity scores under active treatment and control treatment regimes.
  - Plot of superimposed densities of key covariates under active treatment and control treatment regimes.

Weighting the above metrics according to importance of covariates provides yet anther extension.

Further inspection of histograms and estimated propensity scores from each of our imputed datasets (omitted here) confirmed satisfactory covariate balance.

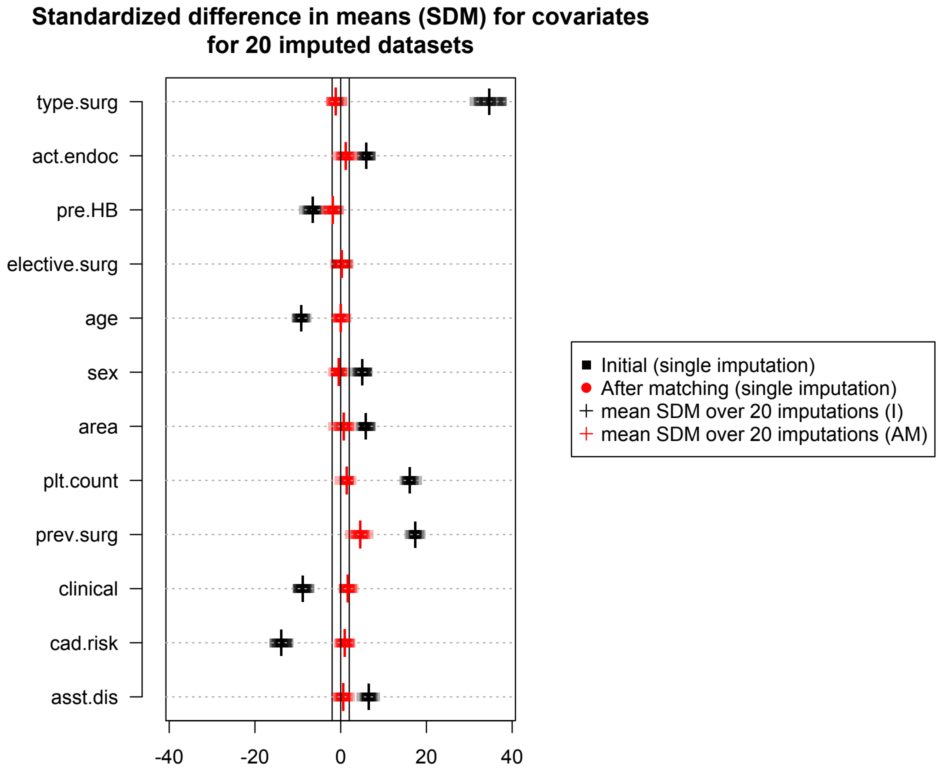

Figure 3.5: *Standardized difference in means, initial and after matching, for covariates, for 20 imputed datasets.*

Note: to ensure fair before and after comparison, post-matching differences in means were standardized using the estimate of the variance of differences in means before matching. Displayed post-matching statistics are thus not t-statistics in the conventional sense.

Vertical lines appear at standardized differences in means of -2, 0 and 2, respectively.

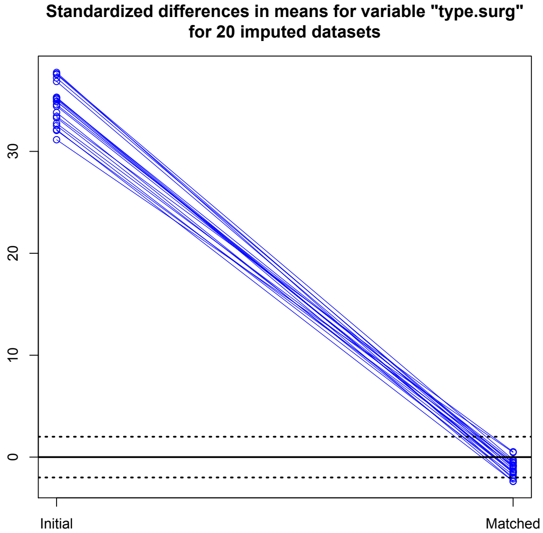

Figure 3.6: *Standardized differences in means for variable type.surg, for 20 imputed datasets.*
Reduced absolute difference in means after matching are represented in blue.

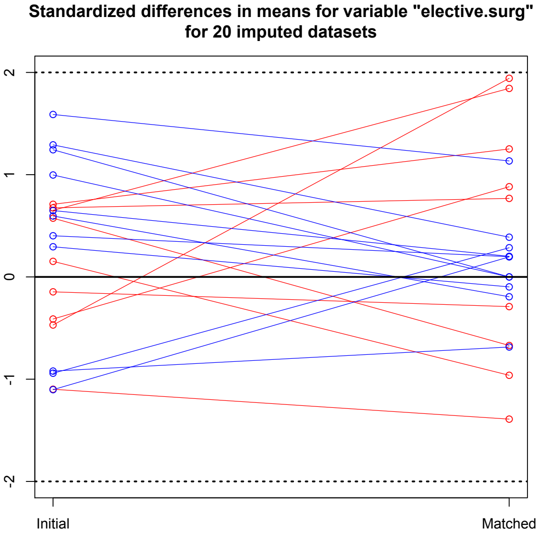

Figure 3.7: *Standardized differences in means for variable elective.surg, for 20 imputed datasets.*
Reduced absolute difference in means after matching are represented in blue. Increased, in red.

3.3.3 Analysis outline and results

We obtain point estimates and 95% confidence intervals for τ_{SP} and ω_{SP} (see Table 3.4) by applying methods from Section 3.2.2 to each imputed realized population cohort dataset, and combining ensuing results using Rubin's rules for multiple imputation (Rubin, 1987).

Table 3.4: *Causal analysis of realized sample cohort dataset via PrepDA: results.*

estimand	**method**	**point estimate**	**95% CI**
risk difference	Neyman	0.013	(-0.025, 0.052)
	Regression	0.005	(-0.033, 0.043)
odds ratio	Woolf	1.20	(0.71, 2.05)
	Regression	1.34	(0.69, 2.62)

Table 3.4 above indicates that all methods yield point estimates and confidence intervals that are in close agreement with population cohort analysis benchmarks (Table 3.3).

3.4 Associative analysis of realized sample cohort dataset via logistic regression

In this section, we use logistic regression to analyze, once again, the realized sample cohort dataset. In addition, we explore the potential for analyst bias with logistic regression methods.

3.4.1 Analysis outline and results

We analyze data according to the following pre-specified protocol. We regress *post-operative renal falure* on the indicator for aprotinin receipt and all twelve covariates. We also fit a model with all main effects and 2-way interactions between covariates. Then, starting with the aforementioned models, we run forward and backward stepwise model selection (Hocking, 1976), using both the AIC (Akaike, 1973) and BIC

(Gideon, 1978) information criteria. Point estimates and 95% confidence intervals for the associative odds ratio, $\omega_{SP|W,\boldsymbol{x}}$, are reported in Table 3.5 below.

Table 3.5: *Associative analysis of realized sample cohort dataset via logistic regression: results.*

baseline model	**model selection method**	**point estimate**	**95% CI**
main effects	none	2.62	(1.21, 5.69)
	forward model selection (AIC)	2.62	(1.21, 5.69)
	backward model selection (AIC)	2.33	(1.16, 4.71)
	forward model selection (BIC)	2.62	(1.21, 5.69)
	backward model selection (BIC)	2.01	(1.01, 3.99)
main effects and 2-way interactions	none	2.96	(1.06, 8.27)
	forward model selection (AIC)	2.96	(1.06, 8.27)
	backward model selection (AIC)	3.10	(1.16, 8.29)
	forward model selection (BIC)	2.96	(1.06, 8.27)
	backward model selection (BIC)†	2.21	(0.96, 5.08)

As can be seen, results generated by the above regression models stand in contrast to population cohort analysis benchmarks (Table 3.3). In particular, none of the estimates are contained within the 95% confidence intervals produced by Woolf's method, and only 3 out of the 10 estimates fall within the regression-generated interval from Section 3.2. What's more, 9 out of 10 outputs suggest a statistically significant relationship between aprotinin intake and postoperative renal failure, which stands in disagreement with benchmark results (as well as with model †, Table 3.5, results).

3.4.2 The perils of logistic regression

Table 3.5 shows that standard model selection procedures alone can produce two sets of results that imply substantively differing study conclusions. So as to further explore the potential for analyst bias, we distort, to the extent possible, regression results by means of deliberate model selection. The outcome of this exercise — i.e., the resulting two models with most contrast results — is presented in Table 3.6. Model details are provided in Appendix B.5.

Table 3.6: *Associative analysis of realized sample cohort dataset via logistic regression, with analyst bias: results.*

model	**point estimate**	**CI**	**Hosmer-Lemeshow test p-value**
1	2.11	(0.89, 4.99)	0.79
2	3.78	(1.36, 10.54)	0.11

The first model indicates a non-significant association between intake of aprotinin versus tranexamic acid, and postoperative renal failure. The second suggests otherwise, and yields an estimate of the odds ratio that is approximately 1.8-fold that generated by model 1. Both models fit the data well per the Hosmer-Lemoshow test, and are arguably reasonable in that they were obtained by simply adding quadratic terms to models from Section 3.4.1 analysis.

The above exercise shows that regression, applied to case-noncase data, provides analysts with ample opportunity to fish for sought-after results.

3.5 Conclusion

We showed that, when applied to sample cohort data, and under the assumption of correct imputation model, our technology can produce estimates for the super-population risk difference and odds ratio that are in agreement with those obtained via application of standard causal inference methods to population cohort data. These findings demonstrate — if only conceptually, in the artificial setting of known imputation model — the potential of PrepDA for the analysis of real-life case-noncase study data.

In contrast, disparities between estimates of the associative odds ratio generated by logistic regression on case-noncase data, and estimates of the causal odds ratio produced by a prospective causal analysis, suggest that regression methods may be inappropriate for purposes of causal inference with case-noncase data. This fact was previously established for prospective designs (see, e.g., Cochran, 1957; Cochran and Rubin, 1973; Rubin, 1973b, 2001).

Last but not least, our study demonstrates the perils of logistic regression with regards to objectivity of (causal) analysis. Our empirical example shows that regression adjustment can be misused, via intentional model construction and selection, on case-noncase data, to produce biased results. PrepDA guards against such practice.

Chapter 4

Discussion and future work

The causal inference framework introduced in Chapter 1 extends the many benefits of the Rubin Causal Model to retrospective studies. Through formulation of the case-noncase study as a cohort study with missing data, our approach fills a conceptual gap between (observational) prospective cohort studies and retrospective studies. Conceptual coherence ensues: a case-noncase study is a partially observed cohort study, which itself is a broken stratified randomized experiment. The problem of causal inference for retrospective studies is therefore conceptually identical to that for cohort studies: the challenge is to reconstruct, to the extent possible, the broken randomized experiment.

Much like the RCM, our approach focuses on first principles: problem definition, framework setup, missing data theory, and Bayesian multiple imputation. This, in turn, allows for clear formulation of assumptions (e.g., unconfoundedness) made in reaching causal conclusions. It also discourages the careless application of standard statistical techniques, such as regression adjustment, for purposes of causal analysis

— the inclusion of outcome variables as predictors in a regression model comes to mind. Last but not least, our approach provides a deeper understanding of the causal inference problem for case-noncase studies than traditional methodologies. For instance, as shown in Section 1.2.9, the potential outcomes perspective sheds new light on the age-old controversy over the use of retrospective matching. The benefits of potential outcomes-based causal inference, and of the missing data perspective in applied and theoretical statistical problems, are further discussed in Rubin (2005a,b).

Our recommendation is that analysts ask a series of key questions when tackling any given causal inference problem with case-noncase data. These include, but are not limited to, "What estimand is of practical relevance?", "How and why is the case-noncase study better suited to addressing the problem?", "Is the assignment mechanism unconfounded?", "Is the population cohort well-defined?", "What scientific knowledge can be incorporated into the study?", and "How should potential outcomes be modeled?".

A fundamental difference between our approach and the framework underlying logistic regression is that the estimand is not forced to be a parameter in some super-population model. As shown in Chapters 1 and 2, our technology allows for the definition, and inference, of *population cohort* (i.e., finite population) causal estimands, and thereby focuses on quantities that we believe are of greater practical interest. In contrast, most epidemiological techniques generally focus on abstract super-population statistical quantities. What's more, our estimands are, by definition, causal. This is in contrast to the oft-studied "associative" odds ratio

$$\frac{\Pr(Y(W)=1|W=1)}{\Pr(Y(W)=0|W=1)} \Bigg/ \frac{\Pr(Y(W)=1|W=0)}{\Pr(Y(W)=0|W=0)} \tag{4.1}$$

(and other standard measures of association) which, as Holland and Rubin (1988) demonstrate, has generally no causal relevance. Lastly, the choice of estimand is flexible under our approach. That is to say, analyses are not restricted to logistic and other multiplicative intercept models (as argued by Wacholder, 1996), nor to the associative odds ratio, which is ubiquitous in epidemiological studies, in part because of its non-sensitivity to choice of sampling design (prospective or retrospective; see, e.g., Bishop et al., 1975). The NSMIT method, introduced in Section 2.2, for instance, can estimate the population cohort risk difference.

PrepDA guarantees objectivity, to the extent we believe is possible, in the estimation of causal effects from case-noncase data. In contrast, as shown in Chapter 3, regression adjustment can be misused, via intentional model construction and selection, to produce biased results. Moreover, our findings suggest that regression methods may be inappropriate for purposes of causal inference with case-noncase data.

From a methodological perspective, our technology allows for the application of the widely-accepted matched sampling methods for pretreatment bias reduction (Rubin, 2006; Imbens and Rubin, 2015) to retrospectively collected data, post-data preprocessing. It also enables, and should encourage, investigators to assess the validity of their causal findings, e.g. by checking for overlap in covariate distributions between treatment groups, also post-preprocessing. A standard pre-specified analysis (e.g., pre-specified regression analysis) of case-noncase data, though objective, does not.

Also, PrepDA can be extended to accommodate secondary outcome analysis, via incorporation of secondary outcomes into the model for population cohort potential outcomes. Future work will explore this topic.

Our simulations show that PrepDA-based methods NSMIT and HGSMIT generally yield reasonable coverage rates, although they under-cover in the presence of covariate imbalance. Also, both methods are effective in reducing bias, except when MI produces accurate estimates of the risk difference and odds ratio. In this type of setting, NSMIT and HGSMIT are prone to introducing minor noise in the overall estimation of causal effects. In our view, this disadvantage is trumped by the considerable benefit of analyst objectivity guaranteed by PrepDA. Moreover, HGSMIT generally yields similar results to LR in our controlled simulation settings. However, the former (and MI) appears to be less sensitive to link function misspecification than the latter (and PR) with regards to bias reduction and mean interval width properties.

A disadvantage of our technology, in its current form, however, is its dependence on party A's choice of imputation model. Accordingly, future research will investigate the use of non-parametric imputation methods in the preprocessing step of PrepDA. The use of spline regression methods, for instance, has been shown effective for the analysis of prospective observational study data (Gutmanan and Rubin, 2012). Future research will also focus on the generalization of NSMIT and HGSMIT to more realistic settings (e.g., in the presence of a mixture of continuous and categorical covariate data). Thus far, our efforts have indicated this to be a difficult undertaking, for both analytical and computational reasons. Lastly, our method is computationally intensive, whereas traditional methods generally are not.

To conclude, our first exploration into bringing objectivity in causal inference with case-noncase data suggests a tradeoff to be had between (a) objectivity of the analysis of case-noncase data, and (b) inferential simplicity, computational efficiency, and — potentially — robustness (i.e., non-reliance on modeling assumptions). Despite the aforementioned shortcomings of PrepDA, the findings of this thesis nevertheless demonstrate our methodology's potential for the objective and causal analysis of real-life case-noncase data.

Appendix A

Supplement to Chapter 2

A.1 PrepDA, step 1: prior specification and posterior predictive draws

A.1.1 Prior specification on $\boldsymbol{\theta}$

$$\begin{aligned}
p &\sim \text{Beta}(2,2) && \text{(A.1)}\\
\mu_0 &\sim \mathcal{N}(0,10) && \text{(A.2)}\\
\sigma_0 &\sim \text{Inv-Gamma}(1.5,2.5) && \text{(A.3)}\\
\mu_1 &\sim \mathcal{N}(0,10) && \text{(A.4)}\\
\sigma_1 &\sim \text{Inv-Gamma}(1.5,2.5) && \text{(A.5)}\\
\beta_0^{(0)} &\sim \text{Cauchy}(0,10) && \text{(A.6)}\\
\beta_X^{(0)} &\sim \text{Cauchy}(0,10) && \text{(A.7)}\\
\beta_0^{(1)} &\sim \text{Cauchy}(0,10) && \text{(A.8)}\\
\beta_X^{(1)} &\sim \text{Cauchy}(0,10) && \text{(A.9)}
\end{aligned}$$

A.1.2 Imputation of $\tilde{\boldsymbol{Y}}^{\mathbf{mis}}$: drawing from $f(\tilde{\boldsymbol{Y}}^{\mathbf{mis}}|\tilde{\boldsymbol{Y}}^{\mathbf{obs}},\boldsymbol{\theta})$

- For those units for which $S_i^{\text{r},\ \cdot} = 1$:

1. If $W_i^{\cdot,\text{ inc}} = 0$:

$$\begin{aligned}
\Pr(Y_i^{\text{mis, inc}} = 1|\tilde{\boldsymbol{Y}}^{\text{obs}}, \boldsymbol{\theta}) &= \Pr(Y_i^{\text{mis, inc}} = 1|Y_i^{\text{r, inc}}, W_i^{\cdot,\text{ inc}} = 0, X_i^{\cdot,\text{ inc}}, S_i^{\text{r},\cdot} = 1, \boldsymbol{\theta}) \\
&= \Pr(Y_i^{\text{mis, }\cdot} = 1|Y_i^{\text{r, }\cdot}, W_i = 0, X_i, S_i^{\text{r, }\cdot} = 1, \boldsymbol{\theta}) \\
&= \Phi(\beta_0^{(1)} + \beta_X^{(1)} X_i) \qquad \text{(A.10)}
\end{aligned}$$

2. If $W_i^{\cdot,\text{ inc}} = 1$:

$$\begin{aligned}
\Pr(Y_i^{\text{mis, inc}} = 1|\tilde{\boldsymbol{Y}}^{\text{obs}}, \boldsymbol{\theta}) &= \Pr(Y_i^{\text{mis, inc}} = 1|Y_i^{\text{r, inc}}, W_i^{\cdot,\text{ inc}} = 0, X_i^{\cdot,\text{ inc}}, S_i^{\text{r},\cdot} = 1, \boldsymbol{\theta}) \\
&= \Pr(Y_i^{\text{mis, }\cdot} = 1|Y_i^{\text{r, }\cdot}, W_i = 0, X_i, S_i^{\text{r, }\cdot} = 1, \boldsymbol{\theta}) \\
&= \Phi(\beta_0^{(0)} + \beta_X^{(0)} X_i) \qquad \text{(A.11)}
\end{aligned}$$

- For those units for which $S_i^{\text{r, }\cdot} = 0$:

  Given $\boldsymbol{\theta}$, independently for each i, draw sequentially from conditional distributions using the following:

  1.

$$\begin{aligned}
\Pr(W_i^{\cdot,\text{ exc}} = 1|\tilde{\boldsymbol{Y}}^{\text{obs}}, \boldsymbol{\theta}) &= \Pr(W_i^{\cdot,\text{ exc}} = 1|S_i^{\text{r},\cdot} = 0, \boldsymbol{\theta}) \qquad \text{(A.12)} \\
&= \frac{p \cdot \left[1 - \Phi\left\{\frac{\beta_0^{(1)} + \beta_X^{(1)}\mu_1}{\sqrt{1 + \sigma_1^2 \beta_X^{(1)^2}}}\right\}\right]}{p \cdot \left[1 - \Phi\left\{\frac{\beta_0^{(1)} + \beta_X^{(1)}\mu_1}{\sqrt{1 + \sigma_1^2 \beta_X^{(1)^2}}}\right\}\right] + (1-p) \cdot \left[1 - \Phi\left\{\frac{\beta_0^{(0)} + \beta_X^{(0)}\mu_0}{\sqrt{1 + \sigma_0^2 \beta_X^{(0)^2}}}\right\}\right]}
\end{aligned}$$

2.

$$
\begin{aligned}
\Pr(X_i^{\cdot,\text{ exc}}|W_i = 0, \tilde{\boldsymbol{Y}}^{\text{obs}}, \boldsymbol{\theta}) &= \Pr(X_i^{\cdot,\text{ exc}}|W_i = 0, S_i^{\text{r},\cdot} = 0, \boldsymbol{\theta}) \\
&= \frac{(1 - \Phi\{\beta_0^{(0)} + \beta_X^{(0)} x_i\}) \cdot \phi(x_i; \mu_0, \sigma_0^2)}{1 - \Phi\Big\{\frac{\beta_0^{(0)} + \beta_X^{(0)} \mu_0}{\sqrt{1 + \sigma_0^2 {\beta_X^{(0)}}^2}}\Big\}} \qquad \text{(A.13)} \\
\Pr(X_i^{\cdot,\text{ exc}}|W_i = 1, \tilde{\boldsymbol{Y}}^{\text{obs}}, \boldsymbol{\theta}) &= \Pr(X_i^{\cdot,\text{ exc}}|W_i = 1, S_i^{\text{r},\cdot} = 1, \boldsymbol{\theta}) \\
&= \frac{(1 - \Phi\{\beta_0^{(1)} + \beta_X^{(1)} x_i\}) \cdot \phi(x_i; \mu_1, \sigma_1^2)}{1 - \Phi\Big\{\frac{\beta_0^{(1)} + \beta_X^{(1)} \mu_1}{\sqrt{1 + \sigma_1^2 {\beta_X^{(1)}}^2}}\Big\}} \qquad \text{(A.14)}
\end{aligned}
$$

We sample from the above two distributions via grid sampling.

3 . Let

$$
\begin{aligned}
P_{(j,k)|w}^{(i)} &:= \Pr((Y_i(0), Y_i(1))^{\text{mis, exc}} = (j,k)|X_i, W_i = w, \tilde{\boldsymbol{Y}}^{\text{obs}}, \boldsymbol{\theta}) \\
&= \Pr((Y_i(0), Y_i(1))^{\text{mis, exc}} = (j,k)|X_i, W_i = w, S_i^{\text{obs}} = 0, \boldsymbol{\theta}) \qquad \text{(A.15)}
\end{aligned}
$$

Then

$$P_{(0,0)|0}^{(i)} = 1 - \Phi(\beta_0^{(1)} + \beta_X^{(1)} X_i) \qquad \text{(A.16)}$$

$$P_{(0,0)|1}^{(i)} = 1 - \Phi(\beta_0^{(0)} + \beta_X^{(0)} X_i) \qquad \text{(A.17)}$$

$$P_{(0,1)|0}^{(i)} = \Phi(\beta_0^{(1)} + \beta_X^{(1)} X_i) \qquad \text{(A.18)}$$

$$P_{(1,0)|1}^{(i)} = \Phi(\beta_0^{(0)} + \beta_X^{(0)} X_i) \qquad \text{(A.19)}$$

$$P_{(0,1)|1}^{(i)} = P_{(1,0)|0}^{(i)} = P_{(1,1)|0}^{(i)} = P_{(1,1)|1}^{(i)} = 0 \qquad \text{(A.20)}$$

A.2 A potential outcomes alternative to Woolf's method for inferring (causal) effects from case-noncase data: a simulation study

A.2.1 The model

We assume the setup outlined in Section 1.2, with covariates suppressed. Let p_0, p_1, ϕ, and $\pi \in (0,1)$. Independently for each unit i, $i = 1, \cdots, N$, in the population cohort, we posit the following models for purposes of population cohort data generation and Bayesian inference:

- Potential outcomes:

$$\begin{aligned}\Pr(Y_i(0) = y_{i0}, Y_i(1) = y_{i1} | p_0, p_1) &= p(Y_i(0) = y_{i0} | p_0) \cdot p(Y_i(1) = y_{i1} | p_1) \\ &= p_0^{y_{i0}}(1-p_0)^{1-y_{i0}} \cdot p_1^{y_{i1}}(1-p_1)^{1-y_{i1}} \quad \text{(A.21)}\end{aligned}$$

- Assignment mechanism:

$$\Pr(W_i = w | Y_i(0), Y_i(1), p_0, p_1, \phi) = \phi^w (1-\phi)^{1-w}, \quad \text{(A.22)}$$

  for $w \in \{0, 1\}$.

- Sampling mechanism:

  We assume sampling of all cases, and simple random sampling with known probability π of noncases, which yields the following:

$$
\begin{aligned}
\Pr(S_i(0) = s_{i0}, S_i(1) = s_{i1} | W_i, Y_i(0), Y_i(1), p_0, p_1, \phi) &= p(S_i(0) = s_{i0} | Y_i(0)) \\
&\times p(S_i(1) = s_{i1} | Y_i(1)) \\
&= \left(\pi^{s_{i0}} (1-\pi)^{(1-s_{i0})}\right)^{\mathbf{1}[Y_i(0)=0]} \\
&\times \left(\pi^{s_{i1}} (1-\pi)^{(1-s_{i1})}\right)^{\mathbf{1}[Y_i(1)=0]}
\end{aligned} \tag{A.23}
$$

A.2.2 Choice of priors

Priors were chosen in accordance with parameter simulation settings. We posit:

$$\phi \sim \text{Beta}(2, 2) \tag{A.24}$$

$$p_0 \sim \text{Beta}\Big(\frac{1}{3}, 5\Big) \tag{A.25}$$

$$p_1 \sim \text{Beta}\Big(\frac{1}{3}, 5\Big) \tag{A.26}$$

A.2.3 Simulation study

The following procedure is implemented:

1. Set parameter values $\{N, p_0, p_1, \phi, \pi\}$, as specified in Table A.1.

2. Repeat, x 1000:

   - Given parameter values, generate population cohort data.

   - Obtain sample cohort data from population cohort data.

- Get posterior draws given observed data, using Bayesian model above and Gibbs sampling. (Note: starting points are sampled from an over-dispersed distribution. The Gelman-Rubin (G-R) statistic is computed to verify convergence.) Obtain super-population point estimates of τ and ω and the corresponding credible intervals by taking the mean and (0.025, 0.975) quantiles, respectively, of the parameters' posterior distributions.
- Multiply impute missing data via draws from the posterior predictive distribution of missing data given observed data. From imputed datasets, calculate the risk differences and odds ratios. Obtain finite-population point estimates of τ and ω and the corresponding intervals by taking the mean and (0.025, 0.975) quantiles, respectively of the two sets of calculated estimands.
- Compute odds ratio estimates and 95% confidence intervals using Woolf's (1955) method for those datasets for which Woolf's estimates exists (i.e., datasets for which all observed data counts n_{ij} are nonzero).

3. Assess frequentist properties of above procedures. Namely, for those datasets to which Woolf's method applies, compute, for all methods:

    - mean coverage (nominal coverage level used: 95%)
    - mean absolute percent bias
    - mean interval width

A.2.4 Simulation results

Table A.1: *Sample cohort data sample size $\bar{n}$, % of datasets invalid for comparison, and G-R statistic (Gibbs).*
[1]‘draw’ refers to drawing from the prior distribution of the parameters.

Parameter values								
N	p_0	p_1	ω	ϕ	π	$\overline{n}$	**% of datasets invalid for comparison**	**G-R statistic**
10,000	.01	.01	1	.5	.01	198	0	1.01
10,000	.005	.005	1	.5	.01	149	0	1.01
10,000	.01	.005	.50	.5	.01	174	0	1.01
10,000	.01	.005	.50	.3	.01	184	0	1.01
10,000	.001	.0005	.50	.5	.01	107	10	1.01
10,000	.001	.0005	.50	.5	.1	1,004	8	1.00
10,000	.1	.01	.09	.5	.01	644	0	1.01
10,000	.01	.009	.90	.5	.01	194	0	1.01
10,000	.005	.01	2.01	.5	.01	173	0	1.01
10,000	.005	.01	2.01	.3	.01	164	0	1.01
10,000	.0005	.001	2.00	.5	.01	107	9	1.01
10,000	.0005	.001	2.00	.5	.1	1,006	9	1.00
10,000	.01	.1	11	.5	.01	645	0	1.01
10,000	.009	.01	1.11	.5	.01	194	0	1.01
10,000	.001	.01	10.1	.5	.01	154	0	1.01
10,000	.0005	.01	20.2	.5	.01	151	8	1.01
10,000	.00005	.001	20.2	.5	.01	106	77	1.01
10,000	*draw*[1]	*draw*	*draw*	*draw*	0.01	731	–	

Table A.2: *Mean coverage of nominal 95% intervals: Bayesian Multiple Imputation vs Woolf's method.*
[1]'draw' refers to drawing from the prior distribution of the parameters.

parameter values						**coverage**				
N	p_0	p_1	ω	ϕ	π	**Bayesian MI**				**Woolf's** (ω)
						finite pop.		**super-pop.**		
						τ	ω	τ	ω	
10,000	.01	.01	1	.5	.01	.95	.95	.94	.94	.95
10,000	.005	.005	1	.5	.01	.94	.94	.94	.94	.94
10,000	.01	.005	.50	.5	.01	.96	.97	.96	.97	.96
10,000	.01	.005	.50	.3	.01	.95	.95	.95	.95	.96
10,000	.001	.0005	.50	.5	.01	.97	.96	.96	.97	.98
10,000	.001	.0005	.50	.5	.1	.95	.95	.96	.96	.98
10,000	.1	.01	.09	.5	.01	.95	.95	.95	.95	.95
10,000	.01	.009	.90	.5	.01	.95	.94	.94	.94	.95
10,000	.005	.01	2.01	.5	.01	.95	.95	.94	.94	.94
10,000	.005	.01	2.01	.3	.01	.94	.94	.95	.95	.95
10,000	.0005	.001	2.00	.5	.01	.96	.96	.95	.97	.98
10,000	.0005	.001	2.00	.5	.1	.95	.96	.96	.96	.98
10,000	.01	.1	11	.5	.01	.95	.95	.95	.95	.95
10,000	.009	.01	1.11	.5	.01	.96	.96	.96	.95	.96
10,000	.001	.01	11	.5	.01	.96	.96	.94	.96	.97
10,000	.0005	.01	20.2	.5	.01	.96	.95	.95	.98	.97
10,000	.00005	.001	20.2	.5	.01	.96	.96	.93	.88	.83
10,000	*draw*[1]	*draw*	*draw*	*draw*	0.01	**.95**	**.95**	.95	.95	–

Table A.3: *Mean absolute percent bias: Bayesian Multiple Imputation vs Woolf's method.*
*Mean absolute percent bias is ill-defined because $\tau = 0$.

parameter values						**mean abs. % bias**				
						Bayesian MI				**Woolf's** (ω)
N	p_0	p_1	ω	ϕ	π	**finite pop.**		**super-pop.**		
						τ	ω	τ	ω	
10,000	.01	.01	1	.5	.01	–*	23.6	–*	24.3	23.4
10,000	.005	.005	1	.5	.01	–*	30.3	–*	32.7	30
10,000	.01	.005	.50	.5	.01	36.5	26.0	36.5	26.9	25.7
10,000	.01	.005	.50	.3	.01	36.6	29.3	35.9	30.2	29.8
10,000	.001	.0005	.50	.5	.01	120.9	88.0	115.1	167.4	86.0
10,000	.001	.0005	.50	.5	.1	130.1	76.7	272	148.3	76.7
10,000	.1	.01	.09	.5	.01	10.0	21.5	10.0	21.6	21.2
10,000	.01	.009	.90	.5	.01	254.1	24.7	99.2	1127.6	1079.1
10,000	.005	.01	2.01	.5	.01	38.5	28.1	38.6	29.9	28.0
10,000	.005	.01	2.01	.3	.01	46.1	29.2	46.0	30.5	29.9
10,000	.0005	.001	2.00	.5	.01	117.7	73.2	116.2	211.6	70.5
10,000	.0005	.001	2.00	.5	.1	112.2	70.7	111.9	206.4	69.1
10,000	.01	.1	11	.5	.01	10.3	21.3	10.3	21.9	21.5
10,000	.009	.01	1.11	.5	.01	221.1	23.7	212.3	23.7	22.8
10,000	.001	.01	11	.5	.01	16.3	60.7	16.2	118.3	58.4
10,000	.0005	.01	20.2	.5	.01	15.7	64.3	15.6	190.1	59.8
10,000	.00005	.001	20.2	.5	.01	51.5	39.5	43.6	41.8	76.3

Table A.4: *Mean width of nominal 95% intervals: Bayesian Multiple Imputation vs Woolf's method.*

parameter values						**mean interval width**				
						Bayesian MI				
N	p_0	p_1	ω	ϕ	π	**finite pop.**		**super-pop.**		**Woolf's** (ω)
						τ	ω	τ	ω	
10,000	.01	.01	1	.5	.01	.01	1.04	.01	1.23	1.24
10,000	.01	.01	1	.5	.01	.01	1.26	.01	1.62	1.62
10,000	.01	.005	.50	.5	.01	.01	.56	.01	.68	.69
10,000	.01	.005	.50	.3	.01	.01	.67	.01	.77	.81
10,000	.001	.0005	.50	.5	.01	.00	2.05	.00	4.66	4.13
10,000	.001	.0005	.50	.5	.1	.00	1.81	.00	4.21	3.72
10,000	.1	.01	.09	.5	.01	.04	.09	.05	.10	.10
10,000	.01	.009	.90	.5	.01	.01	.97	.01	1.15	1.16
10,000	.005	.01	2.01	.5	.01	.01	2.31	.01	2.91	2.89
10,000	.005	.01	2.01	.3	.01	.01	2.46	.01	2.98	3.08
10,000	.0005	.001	2.00	.5	.01	.00	7.05	.00	24.95	16.65
10,000	.0005	.001	2.00	.5	.1	.00	6.40	.00	23.57	15.47
10,000	.01	.1	11	.5	.01	.04	10.64	.05	12.01	12.25
10,000	.009	.01	1.11	.5	.01	.01	1.15	.01	1.36	1.37
10,000	.001	.01	11	.5	.01	.01	23.92	.01	63.7	45.75
10,000	.0005	.01	20.2	.5	.01	.01	52.36	.01	220.6	133.12
10,000	.00005	.001	20.2	.5	.01	.00	14.37	.00	68.34	40.45

A.3 Example prior specification, and posterior predictive draws for MVN extension model

A.3.1 Prior specification on $\boldsymbol{\theta}$

$$\begin{aligned}
p &\sim \text{Beta}(2, 2) && \text{(A.27)}\\
\boldsymbol{\mu}_0 &\sim \mathcal{MVN}_k(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) && \text{(A.28)}\\
\boldsymbol{\Sigma}_0 &\sim \text{Inv-Wishart}(k+1, \boldsymbol{\Sigma}^\circ) && \text{(A.29)}\\
\boldsymbol{\mu}_1 &\sim \mathcal{MVN}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) && \text{(A.30)}\\
\boldsymbol{\Sigma}_1 &\sim \text{Inv-Wishart}(k+1, \boldsymbol{\Sigma}^\circ) && \text{(A.31)}\\
\boldsymbol{\beta}^{(0)} &\sim \mathcal{MVN}_{(k+1)}(\boldsymbol{\mu}^\dagger, \boldsymbol{\Sigma}^\dagger) && \text{(A.32)}\\
\boldsymbol{\beta}^{(1)} &\sim \mathcal{MVN}_{(k+1)}(\boldsymbol{\mu}^\dagger, \boldsymbol{\Sigma}^\dagger) && \text{(A.33)}\\
& && \text{(A.34)}
\end{aligned}$$

where

$$\boldsymbol{\mu}^{*} = (0, \cdots, 0)^{T} \tag{A.35}$$

$$\boldsymbol{\Sigma}^{*} = \begin{vmatrix} 1000 & 0 & \dots & 0 \\ 0 & 1000 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1000 \end{vmatrix} \tag{A.36}$$

$$\boldsymbol{\Sigma}^{\circ} = \begin{vmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{vmatrix} \tag{A.37}$$

$$\boldsymbol{\mu}^{\dagger} = (0, \cdots, 0)^{T} \tag{A.38}$$

$$\boldsymbol{\Sigma}^{\dagger} = \begin{vmatrix} 5 & 0 & \dots & 0 \\ 0 & 5 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 5 \end{vmatrix} \tag{A.39}$$

$$\tag{A.40}$$

The Inverse Wishart prior above induces uniform marginal distributions for all individual correlations (Barnard et al., 2000).

A.3.2 Imputation of $\tilde{\boldsymbol{Y}}^{\mathbf{mis}}$: drawing from $f(\tilde{\boldsymbol{Y}}^{\mathbf{mis}}|\tilde{\boldsymbol{Y}}^{\mathbf{obs}}, \boldsymbol{\theta})$

- For those units for which $S_i^{\mathrm{r},\cdot} = 1$:

1. If $W_i^{\cdot, \text{ inc}} = 0$:

$$\begin{aligned}
\Pr(Y_i^{\text{mis, inc}} = 1 | \tilde{\boldsymbol{Y}}^{\text{obs}}, \boldsymbol{\theta}) &= \Pr(Y_i^{\text{mis, inc}} = 1 | Y_i^{\text{r, inc}}, W_i^{\cdot, \text{ inc}} = 0, \mathbf{x}_i^{\cdot, \text{ inc}}, S_i^{\text{r},\cdot} = 1, \boldsymbol{\theta}) \\
&= \Pr(Y_i^{\text{mis, }\cdot} = 1 | Y_i^{\text{r, }\cdot}, W_i = 0, \mathbf{x}_i, S_i^{\text{r, }\cdot} = 1, \boldsymbol{\theta}) \\
&= \Phi(\widetilde{\mathbf{x}}_i \boldsymbol{\beta}^{(1)}) \qquad \text{(A.41)}
\end{aligned}$$

2. If $W_i^{\cdot, \text{ inc}} = 1$:

$$\begin{aligned}
\Pr(Y_i^{\text{mis, inc}} = 1 | \tilde{\boldsymbol{Y}}^{\text{obs}}, \boldsymbol{\theta}) &= \Pr(Y_i^{\text{mis, inc}} = 1 | Y_i^{\text{r, inc}}, W_i^{\cdot, \text{ inc}} = 0, \mathbf{x}_i^{\cdot, \text{ inc}}, S_i^{\text{r},\cdot} = 1, \boldsymbol{\theta}) \\
&= \Pr(Y_i^{\text{mis, }\cdot} = 1 | Y_i^{\text{r, }\cdot}, W_i = 0, \mathbf{x}_i, S_i^{\text{r, }\cdot} = 1, \boldsymbol{\theta}) \\
&= \Phi(\widetilde{\mathbf{x}}_i \boldsymbol{\beta}^{(0)}) \qquad \text{(A.42)}
\end{aligned}$$

- For those units for which $S_i^{\text{r, }\cdot} = 0$:

  Given $\boldsymbol{\theta}$, independently for each i, draw sequentially from conditional distributions using the following:

  1.

$$\begin{aligned}
\Pr(W_i^{\cdot, \text{ exc}} = 1 | \tilde{\boldsymbol{Y}}^{\text{obs}}, \boldsymbol{\theta}) &= \Pr(W_i^{\cdot, \text{ exc}} = 1 | S_i^{\text{r},\cdot} = 0, \boldsymbol{\theta}) \\
&= \frac{p \cdot \left[1 - \Phi\left\{\frac{\widetilde{\boldsymbol{\mu}}_1^T \boldsymbol{\beta}^{(1)}}{\sqrt{1+\boldsymbol{\beta}^{(1)T} \widetilde{\boldsymbol{\Sigma}}_1 \boldsymbol{\beta}^{(1)}}}\right\}\right]}{p \cdot \left[1 - \Phi\left\{\frac{\widetilde{\boldsymbol{\mu}}_1^T \boldsymbol{\beta}^{(1)}}{\sqrt{1+\boldsymbol{\beta}^{(1)T} \widetilde{\boldsymbol{\Sigma}}_1 \boldsymbol{\beta}^{(1)}}}\right\}\right] + (1-p) \cdot \left[1 - \Phi\left\{\frac{\widetilde{\boldsymbol{\mu}}_0^T \boldsymbol{\beta}^{(0)}}{\sqrt{1+\boldsymbol{\beta}^{(0)T} \widetilde{\boldsymbol{\Sigma}}_0 \boldsymbol{\beta}^{(0)}}}\right\}\right]}
\end{aligned} \qquad \text{(A.43)}$$

2.

$$\Pr(\boldsymbol{x}_i^{\cdot,\text{ exc}}|W_i = 0, \tilde{\boldsymbol{Y}}^{\text{obs}}, \boldsymbol{\theta}) = \Pr(\boldsymbol{x}_i^{\cdot,\text{ exc}}|W_i = 0, S_i^{\text{r},\cdot} = 0, \boldsymbol{\theta}) \tag{A.44}$$

$$= \frac{(1 - \Phi\{\widetilde{\boldsymbol{x}}_i\boldsymbol{\beta}^{(0)}\}) \cdot \phi_k(\boldsymbol{x}_i; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)}{1 - \Phi\left\{\frac{\widetilde{\boldsymbol{\mu}}_0^T\boldsymbol{\beta}^{(0)}}{\sqrt{1+\boldsymbol{\beta}^{(0)\,T}\widetilde{\boldsymbol{\Sigma}}_0\boldsymbol{\beta}^{(0)}}}\right\}} \tag{A.45}$$

$$\Pr(\boldsymbol{x}_i^{\cdot,\text{ exc}}|W_i = 1, \tilde{\boldsymbol{Y}}^{\text{obs}}, \boldsymbol{\theta}) = \Pr(\boldsymbol{x}_i^{\cdot,\text{ exc}}|W_i = 1, S_i^{\text{r},\cdot} = 1, \boldsymbol{\theta})$$

$$= \frac{(1 - \Phi\{\widetilde{\boldsymbol{x}}_i\boldsymbol{\beta}^{(1)}\}) \cdot \phi_k(\boldsymbol{x}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)}{1 - \Phi\left\{\frac{\widetilde{\boldsymbol{\mu}}_1^T\boldsymbol{\beta}^{(1)}}{\sqrt{1+\boldsymbol{\beta}^{(1)\,T}\widetilde{\boldsymbol{\Sigma}}_1\boldsymbol{\beta}^{(1)}}}\right\}}$$

We sample from the above two distributions via STAN.

3 . Let

$$P_{(j,k)|w}^{(i)} := \Pr((Y_i(0), Y_i(1))^{\text{mis, exc}} = (j,k)|\boldsymbol{x}_i, W_i = w, \tilde{\boldsymbol{Y}}^{\text{obs}}, \boldsymbol{\theta})$$

$$= \Pr((Y_i(0), Y_i(1))^{\text{mis, exc}} = (j,k)|\boldsymbol{x}_i, W_i = w, S_i^{\text{obs}} = 0, \boldsymbol{\theta}) \tag{A.46}$$

Then

$$P_{(0,0)|0}^{(i)} = 1 - \Phi(\widetilde{\boldsymbol{x}}_i\boldsymbol{\beta}^{(1)}) \tag{A.47}$$

$$P_{(0,0)|1}^{(i)} = 1 - \Phi(\widetilde{\boldsymbol{x}}_i\boldsymbol{\beta}^{(0)}) \tag{A.48}$$

$$P_{(0,1)|0}^{(i)} = \Phi(\widetilde{\boldsymbol{x}}_i\boldsymbol{\beta}^{(1)}) \tag{A.49}$$

$$P_{(1,0)|1}^{(i)} = \Phi(\widetilde{\boldsymbol{x}}_i\boldsymbol{\beta}^{(0)}) \tag{A.50}$$

$$P_{(0,1)|1}^{(i)} = P_{(1,0)|0}^{(i)} = P_{(1,1)|0}^{(i)} = P_{(1,1)|1}^{(i)} = 0 \tag{A.51}$$

Appendix B

Supplement to Chapter 3

B.1 Difference in covariate means between control and active treatment subgroups

Table B.1: *T-test for difference in covariate means between the two treatment groups, post-matching (Section 3.2 analysis)*
But for the variable *prev.surg*, t-tests suggest non-significant differences in covariate means between the two treatment groups, at the 0.05 significance level, post-matching.

Covariate	**t statistic**	**p-value**
type.surg	-1.15	0.25
act.endoc	0.87	0.39
pre.HB	-1.62	0.10
elective.surg	0.50	0.62
age	-0.23	0.82
sex	-0.22	0.83
area	0.06	0.95
plt.count	1.03	0.30
prev.surg	2.94	0.00
clinical	0.87	0.38
cad.risk	0.35	0.73
asst.dis	0.12	0.90

B.2 Additional histograms

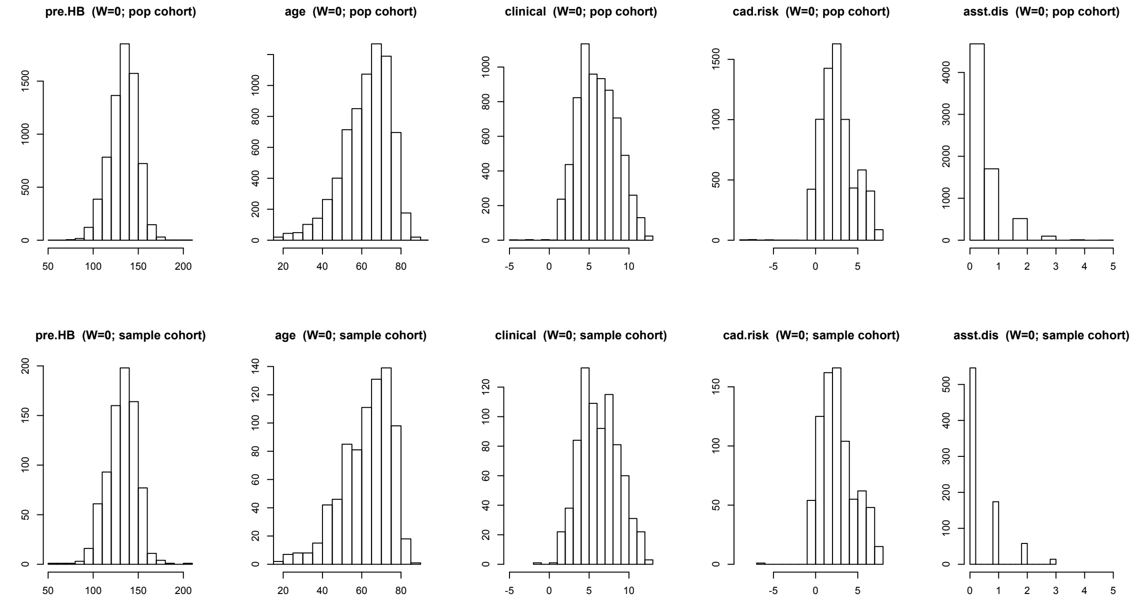

Figure B.1: *Histograms of selected covariates for control treatment group units, displayed by dataset.*

Covariate distributions generally differ between population cohort units and sample cohort units, as expected.

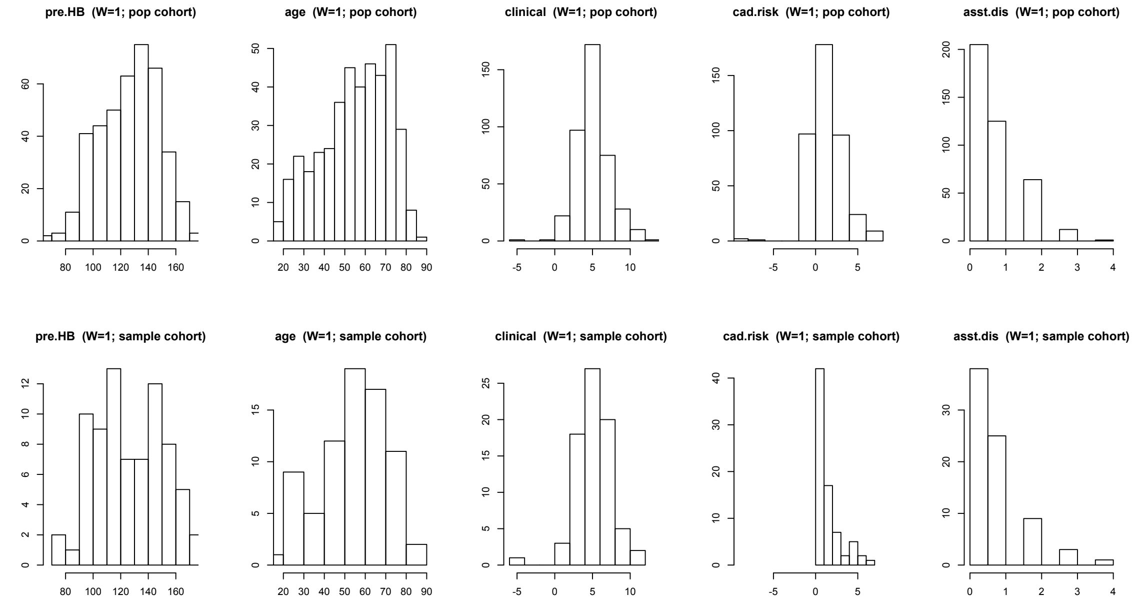

Figure B.2: *Histograms of selected covariates for active treatment group units, displayed by dataset.*
Covariate distributions generally differ between population cohort units and sample cohort units, as expected.

B.3 Zoomed plot of SDM for covariates for 20 imputed datasets

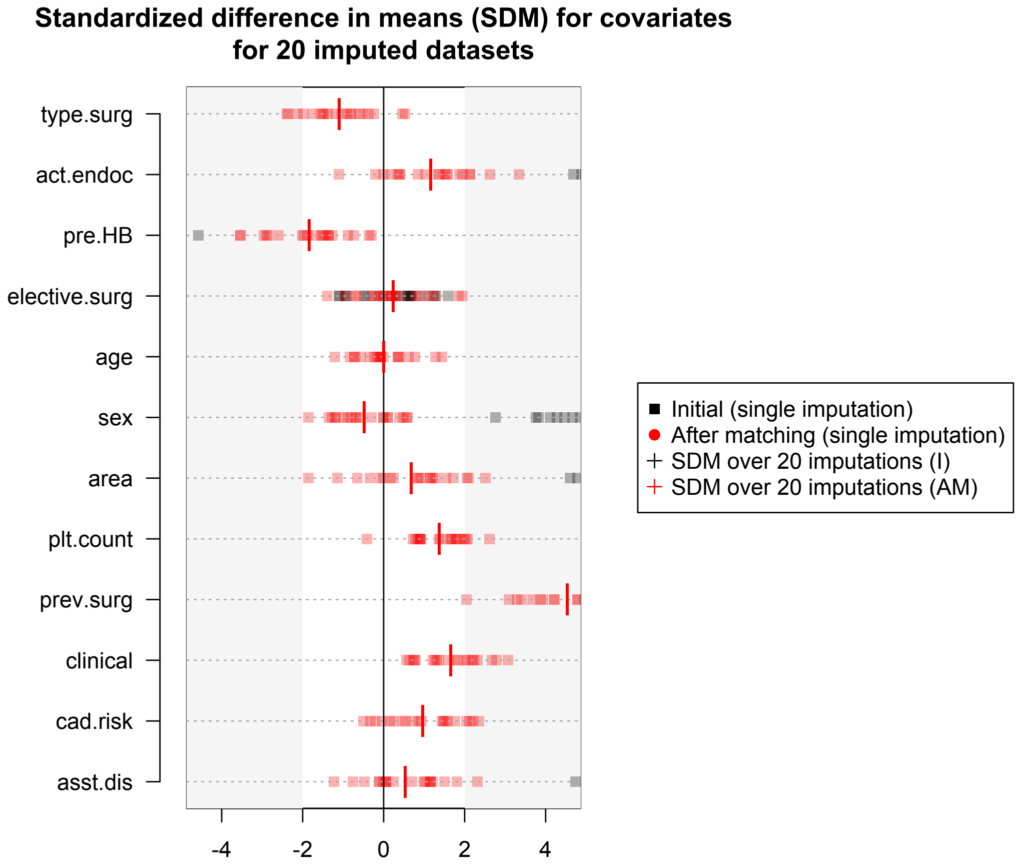

Figure B.3: *Standardized difference in means, initial and after matching, for covariates, for 20 imputed datasets.*
Note that, to ensure fair before-after comparison, we standardize differences in means after matching, using the estimate of the variance of differences in means before matching. As such, displayed after matching statistics are not t-statistics in the conventional sense.
Vertical lines appear at standardized differences in means of -2, 0 and 2, respectively.

B.4 PrepDA, step 1: computational details

We exploit the fact that, within the confines of our study, population cohort data is known. We thus approximate $\Pr(\tilde{\boldsymbol{Y}}^{\text{mis}}|\tilde{\boldsymbol{Y}}^{\text{obs}})$ with the empirical distribution of $\Pr(\tilde{\boldsymbol{Y}}^{\text{mis}}|\tilde{\boldsymbol{Y}}^{\text{r, +}})$, where $\tilde{\boldsymbol{Y}}^{\text{r, +}} = (\boldsymbol{S}^{\text{r, }\cdot}, \boldsymbol{Y}^{\text{r, }\cdot}, \boldsymbol{W}, \boldsymbol{X})$ is the realized population cohort data matrix, supplemented with the vector of realized sampling indicators, $\boldsymbol{S}^{\text{r, }\cdot}$. To sample from $\Pr(\tilde{\boldsymbol{Y}}^{\text{mis}}|\tilde{\boldsymbol{Y}}^{\text{obs}})$, we sample, with replacement, $N - n_{\text{inc}}$ rows

$$\boldsymbol{R}_i^{\circ} := (Y_i^{\text{r, }\cdot}, W_i^{\cdot\text{, inc}}, \boldsymbol{x}_i)\big|_{Y_i^{\text{r, inc}}=0}, \; i \in 1, \cdots, N,$$

from the realized population cohort data matrix.

A singly imputed realized population cohort dataset is obtained by appending the set of sampled rows to the realized sample cohort dataset.

B.5 Specification of logistic regression models

Model 1 was obtained by adding $plt.count^2$, $asst.dis^2$ and $prev.surg^2$ to the model resulting from backward model selection, starting with a model with all main effects and 2-way interaction terms, and using the BIC selection criteria. Its equation is as follows:

$$\begin{aligned} Y \sim W\ +\ type.surg\ +\ pre.HB\ +\ age\ +\ sex\ \ area\ +\ plt.count \\ +prev.surg\ +clinical\ +\ cad.risk\ +\ asst.dis\ +\ type.surg:area \\ +pre.HB:age\ +\ pre.HB:sex\ +pre.HB:asst.dis\ +\ age:area \\ +sex:area\ +\ area:plt.count\ +\ area:prev.surg+area:clinical \end{aligned}$$

Model 2 was obtained by adding $clinical^2$, $type.surg^2$, $prev.surg^2$, $log(age)$ and age^2 to the model resulting from backward model selection, starting with a model with all main effects and 2-way interaction terms, and using the AIC selection criteria. Its equation is as follows:

$$
\begin{aligned}
Y \sim W \; &+ \; \mathit{type.surg} \; + \; \mathit{act.endoc} \; + \; \mathit{pre.HB} \; + \; \mathit{elective.surg} \; + \; \mathit{age} \; + \; \mathit{sex} \\
&+ \; \mathit{area} + \mathit{plt.count} \; + \; \mathit{prev.surg} \; + \; \mathit{clinical} \; + \; \mathit{cad.risk} \; + \; \mathit{asst.dis} \\
&+ \; \mathit{type.surg} : \mathit{elective.surg} + \mathit{type.surg} : \mathit{sex} \; + \; \mathit{type.surg} : \mathit{area} \\
&+ \; \mathit{type.surg} : \mathit{prev.surg} \; + \; \mathit{type.surg} : \mathit{cad.risk} \; + \; \mathit{type.surg} : \mathit{asst.dis} \\
&+ \; \mathit{act.endoc} : \mathit{age} \; + \; \mathit{act.endoc} : \mathit{plt.count} \; + \; \mathit{act.endoc} : \mathit{asst.dis} \\
&+ \mathit{pre.HB} : \mathit{age} \; + \; \mathit{pre.HB} : \mathit{sex} \; + \; \mathit{pre.HB} : \mathit{cad.risk} \; + \; \mathit{pre.HB} : \mathit{asst.dis} \\
&+ \mathit{elective.surg} : \mathit{prev.surg} \; + \; \mathit{elective.surg} : \mathit{asst.dis} \; + \; \mathit{age} : \mathit{area} \\
&+ \mathit{age} : \mathit{prev.surg} \; + \; \mathit{age} : \mathit{cad.risk} \; + \; \mathit{sex} : \mathit{area} \; + \; \mathit{sex} : \mathit{cad.risk} \\
&+ \; \mathit{area} : \mathit{plt.count} \; + \; \mathit{area} : \mathit{prev.surg} \\
&+ \mathit{area} : \mathit{clinical} \; + \; \mathit{plt.count} : \mathit{clinical} \; + \; \mathit{prev.surg} : \mathit{asst.dis} \\
&+ \; \mathit{cad.risk} : \mathit{asst.dis}
\end{aligned}
\tag{B.1}
$$

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second international symposium on information theory*, 267–281.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika 59.*

Anderson, J. A. (1973). *Logistic discrimination with medical applications.* Academic Press.

Barnard, J., R. McCulloch, and X. L. Meng (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica 10*, 1281–1311.

Barry, M. A. et al. (2006). Update: Fusarium keratitis — united states, 2005–2006. `http://www.cdc.gov/mmwr/preview/mmwrhtml/mm55d519a1.htm`. [Online; accessed 20-January-2015].

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge: MIT Press.

Breslow, N. and W. Powers (1978). Are there two logistic regressions for retrospective studies? *Biometrics 34.*

Breslow, N. E. and N. E. Day (1980). *Statistical Methods in Cancer Research. Vol 1: The analysis of case-control studies.* International Agency for Research on Cancer, Lyon, France.

Chretien, Y. (2010). *Three Applications of Statistics to Medical Research.* Ph. D. thesis, Harvard University.

Cochran, W. G. (1957). Analysis of covariance: its nature and uses. *Biometrics 13*, 261 – 281.

Cochran, W. G. and D. B. Rubin (1973). Controlling bias in observational studies: a review. *Sankhyā: The Indian Journal of Statistics, Series A 35*, 417–466.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute 11*, 1269 – 1275.

Costanza, M. C. (1995). Matching. *Preventive Medicine 24*, 425 – 433.

Dehejia, R. H. and S. Wahba (2002). Propensity score-matching methods for non-experimental causal studies. *The Review of Economics and Statistics 84*, 151 – 161.

Ding, P. and T. Dasgupta (2015). A potential tale of two by two tables from completely randomized experiments. *Journal of the American Statistical Association*, in press.

Gallagher, J. (2014). Autism link to air pollution raised. `http://www.bbc.com/news/health-30521255`. [Online; accessed 20-January-2015].

Gart, J. J. (1966). Alternative analyses of contingency tables. *Journal of the Royal Statistical Society Series B 28*, 164–179.

Gefeller, O., A. Pfahlberg, H. Brenner, and J. Windeler (1998). An empirical investigation on matching in published case-control studies. *European Journal of Epidemiology 14*, 321–325.

Gideon, S. (1978). Estimating the dimension of a model. *Second international symposium on information theory 6*, 461–464.

Greenberg, R., S. Daniels, W. Flanders, J. Eley, and J. Boring (2004). *Medical Epidemiology, Edition 4*. McGraw-Hill Professional Publishing.

Greiner, J. (2008). Causal inference in civil rights litigation. *Harvard Law Review 122*, 533 – 598.

Gutmanan, R. and D. B. Rubin (2012). Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes. *Statistics in Medicine 32*, 1795–1814.

Haldane, J. B. S. (1955). The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics 20*, 309 – 311.

Heckman, J. J., H. Hidehiko, and P. Todd (1997). Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *Review of Economic Studies 64*, 605 – 654.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics 32*, 461–464.

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

Holland, P. W. and D. B. Rubin (1988). Causal inference in retrospective studies. *Evaluation Review 12*, 203 – 231.

Imbens, W. I. and D. B. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

Karkouti, K. et al. (2006). A propensity score case-control comparison of aprotinin and tranexamic acid in high-transfusion-risk cardiac surgery. *Transfusion 46*, 327 – 338.

Kousser, J. (1984). Are expert witnesses whores? reflections on objectivity in scholarship and expert witnessing. *The Public Historian 6*, 1–19.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review 76*, 604 – 620.

Lane-Claypon, J. (1926). A further report on cancer of the breast with special reference to its associated antecedent conditions. In *Ministry of Health. Reports on Public Health and Medical Subjects.*

Linden, M. D. (2003). The hemostatic defect of cardiopulmonary bypass. *Journal of Thrombosis and Thrombolysis 16*, 129 – 147.

Little, J. A. R. and D. B. Rubin (2002). *Statistical Analysis with Missing Data.* John Wiley & Sons.

Little, R. (2006). Calibrated bayes: A bayes/frequentist roadmap. *The American Statistician 3*, 213 — 223.

Mantel, N. (1973). Synthetic retrospective studies and related topics. *Biometrics 29.*

Mantel, N. and W. Haenszel (1959). Statistical aspects of the analyses of data from retrospective studies of disease. *Journal of the National Cancer Institute 22*, 719 – 748.

Mayo Clinic Staff (2014a). Endocarditis definition. `http://www.mayoclinic.org/diseases-conditions/endocarditis/basics/definition/con-20022403`. [Online; accessed 08-March-2015].

Mayo Clinic Staff (2014b). Platelet definition. `http://www.mayoclinic.org/diseases-conditions/thrombocytopenia/basics/definition/con-20027170`. [Online; accessed 08-March-2015].

Medicine Net Staff (2014). Dialysis definition. `http://www.medicinenet.com/dialysis/article.htm#1whatis`. [Online; accessed 23-March-2015].

Meng, X. L. (1977). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science =9*, 538 – 558.

Miettinen, O. S. (1970). Estimation of relative risk from individually matched series. *Biometrics 26*, 75 – 86.

Mosby (2009). *Mosby's Dictionary of Medicine, Nursing & Health Professions, 8th Edition*. Elsevier Health Sciences.

Månsson, R. et al. (2007). On the estimation and use of propensity scores in case-control and case-cohort studies. *American Journal of Epidemiology 166*, 332 – 339.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. *Translated in Statistical Science 5*, 465 – 480.

Pattanayak, C. W., D. B. Rubin, and E. R. Zell (2011). Propensity score methods for creating covariate balance in observational studies. *Revista Española de Cardiologa (English Edition) 64*, 897 – 903.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika 73*, 1 – 11.

Prentice, R. L. and R. Pyke (1979). Logistic disease incidence models and case-control studies. *Biometrika 66*, 403 – 11.

Raz, R., A. Roberts, K. Lyall, J. Hart, A. Just, F. Laden, and M. Weisskopf (2014). Autism spectrum disorder and particulate matter air pollution before, during, and after pregnancy: A nested casecontrol analysis within the nurses' health study ii cohort. *Environmental Health Perspectives*.

Robertson, C. T. (2010). Blind expertise. *New York University Law Review 85*, 174 – 257.

Robins, J. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*, 122 – 129.

Rose, S. and M. J. van der Laan (2009a). Causal inference for nested case-control studies using targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series*.

Rose, S. and M. J. van der Laan (2009b). Why match? investigating matched case-control study designs with causal effect estimation. *The International Journal of Biostatistics 5*.

Rothman, K. J., S. Greenland, and T. L. Lash (2008). *Modern Epidemiology.* Lippincott Williams & Wilkins.

Rubin, D. and E. A. Stuart (2007). Matching methods for causal inference: Designing observational studies. *Best Practices in Quantitative Methods (to appear).*

Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics 29*, 159 – 184.

Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics 29*, 185 – 203.

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

Rubin, D. B. (1975). Bayesian inference for causality: The importance of randomization. *The Proceedings of the Social Statistics Section of the American Statistical Association*, 233 – 239.

Rubin, D. B. (1977). Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics 2*, 1 – 26.

Rubin, D. B. (1978a). Bayesian inference for causal effects. *Annals of Statistics 6*, 34 – 58.

Rubin, D. B. (1978b). Multiple imputations in sample surveys – a phenomenological bayesian approach to nonresponse. *The Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20 – 34.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association 74*, 318 – 328.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services & Outcomes Research Methodology 2*, 169 – 188.

Rubin, D. B. (2005a). Causal inference using potential outcomes. *Journal of the American Statistical Association 100*, 322 – 331.

Rubin, D. B. (2005b). Conceptual, computational and inferential benefits of the missing data perspective in applied and theoretical statistical problems. *Allgemeines Statistisches Archiv 90*.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine 26*, 20 – 36.

Rubin, D. B. (2008). Causal inference using potential outcomes. *The Annals of Applied Statistics 2*, 808 – 840.

Rubin, D. B. and N. Thomas (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association 95*, 573 – 585.

Schlesselman, J. J. (1982). *Case-Control Studies: Design, Conduct, Analysis*. Oxford University Press.

Schultz, D. G. (2006). Fda public health notification: Fungal keratitis infections related to contact lens use. `http://www.fda.gov/MedicalDevices/Safety/AlertsandNotices/PublicHealthNotifications/ucm062098.htm`. [Online; accessed 20-January-2015].

Seigel, D. and S. Greenhouse (1973). Multiple relative risk functions in case-control studies. *American journal of epidemiology 97*.

Stan Development Team (2014). Rstan: the r interface to stan, version 2.5.0.

USA Today (2009). Bausch & lomb settles 600 eye fungus lawsuits. `http://usatoday30.usatoday.com/money/industries/health/2009-05-31-bausch-lawsuits_N.htm`. [Online; accessed 20-January-2015].

Wacholder, S. (1996). The case-control study as data missing by design: Estimating risk differences. *Epidemiology 7*, 144 – 150.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics 19*, 251 – 253.