Multiple Imputation of Missing Covariates in Randomized Controlled Trials

by

Gauri Kamat

Department of Statistical Science
Duke University

Date: ____________________________
Approved:

____
Jerome Reiter, Supervisor

____
Fan Li

____
Amy Herring

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2019

Abstract

Multiple Imputation of Missing Covariates in Randomized Controlled Trials

by

Gauri Kamat

Department of Statistical Science
Duke University

Date: ______________________

Approved:

____
Jerome Reiter, Supervisor

____
Fan Li

____
Amy Herring

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistical Science in the Graduate School of Duke University
2019

Abstract

Baseline covariates in randomized experiments play a crucial role in the estimation of treatment effects. Random assignment ensures independence of the covariates and the treatment, which is essential for objective interpretation of the effects. Covariates are also measured before observing the outcome, which guarantees validity of any causal conclusions obtained. When covariates are partly missing, it may be essential to consider if imputations should be carried out respecting randomization or separately by treatment groups. One other question that could arise is if outcome information should be incorporated in the imputations. In view of these considerations, we examine four different ways of multiply imputing missing baseline data in randomized trials. We consider imputation in the design and outcome stages, taking into account, the independence restrictions implied by randomization. We allow for the possibility of non-ignorable missingness, and use identifying restrictions in the nonparametric saturated class to obtain the full data density from the observed data distribution. We further conduct repeated sampling studies to assess the performance of the methods in three different missingness scenarios that could commonly emerge in randomized trials.

To my parents

Contents

List of Tables

List of Figures

Acknowledgements

A big thank you to my supervisor, Jerry Reiter, for his time and patience over the past year. His guidance has played a very important role in shaping this thesis, which would be far worse otherwise. He is an incredible mentor - always encouraging, positive and helpful. I also thank him for all the help with editing and revising thesis drafts.

Thanks to Amy Herring, for some great advice on research and non-research matters. Thanks to Fan Li, for introducing me to Causal Inference and also for being a part of my committee. Many thanks to Surya Tokdar, my academic advisor, who always had genuine advice for me regarding courses and doctoral applications. Special thanks to my family, which has been a constant source of support during my time at Duke.

I also gratefully acknowledge support from the National Science Foundation (SES-11-31897).

1

Introduction

Classical randomized experiments are widely considered to be the benchmark for obtaining unbiased causal effects. With an adequate sample size, random assignment ensures that baseline characteristics of the treatment and control groups do not vary in any systematic manner from each other. Alternatively stated, randomization attains balance on all pretreatment covariates (observed and unobserved), in expectation (Rubin, 2008). Thus, causal comparisons can be made simply by contrasting the outcomes in the two groups. Another appealing feature of randomized trials stems from the two-stage regime usually followed to implement them. For each unit, measurement of covariates and assignment to a treatment group takes place prior to looking at the outcome. This ensures that by design, results do not favour the treatment or the control (Rubin, 2007, 2008).

Often, baseline variables measured in randomized experiments are used in the estimation of causal effects. One example is model-based covariate adjustment, carried out to improve the precision and statistical power of the analysis. Another quantity of interest is the efficacy of the treatment within categories defined by one or more covariates, also called the subgroup causal effect. If the covariates are fully observed, such effects are readily calculated; however, analysis is frequently plagued by missing data. An ad-hoc approach like complete case analysis sacrifices information, which is not only inefficient, but also violates the intention to treat (ITT) principle in randomized studies (Schulz and Grimes, 2002; White and Thompson, 2005).

Following Rubin (1976), we can classify the data to be missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). If the data are MCAR, the probability that a value is missing is unrelated to the observed or unobserved data. If the data are MAR, the probability of missingness is independent of the missing data conditional on the observed data. MNAR entails that this probability depends on the missing values in the dataset. We say that the missingness mechanism is ignorable (Rubin, 1976; Little and Rubin, 2002), if MAR holds and the parameters of the the complete data and the missingness process are distinct, in which case the missingness does not have to be modeled. If the mechanism is non-ignorable, inferences have to be based on the joint distribution of the variables and their missingness indicators. This distribution is under-identified by construction, and inference cannot proceed unless additional assumptions about the missingness mechanism are made. Such assumptions, called identifying restrictions, provide a useful framework for sensitivity analysis (Thijs et al., 2002; Linero & Daniels, 2018).

One approach to handling incomplete data is imputation, or "filling in" missing observations. Single imputation replaces each missing data point by a specific value, say the mean or the mode of the observed data. Although computationally simple, analysis of singly imputed data treats every imputed value as if it were actual data. This fails to incorporate the uncertainty due to missingness, because of which, standard errors are often underestimated. Multiple imputation (MI) (Rubin, 1987) addresses this issue by generating several imputations per missing observation, hence creating multiple completed data sets. Point estimates and variances are calculated from each of these data sets and then combined using the rules outlined in Rubin (1987). Several techniques of implementing MI exist; for a general review, see Harel and Zhou (2007).

Missing outcome data in randomized trials has been discussed extensively in the literature (e.g., Frangakis and Rubin, 1999; Chen et al., 2009; Imai, 2009). Missingness in baseline covariates, however, raises different issues. Two questions arise here. First, should imputations respect the trial protocol and not use the outcome or any other post-randomization measure? Second, should treatment assignment play a role in the imputation, given the marginal independence of the covariates and the treatment? In common regression models, using the outcome to handle missing covariates has been strongly recommended (Rubin & Schencker, 1991; Little, 1992; Vach, 1994; Moons et al., 2006). In the context of randomized experiments, this disturbs the basis that makes them the gold standard for causal evaluations, namely, complete isolation of the design stage from the outcome stage. Accounting for treatment group membership in imputations ignores randomization and may inject incidental imbalances in the observed data into the missing data (White and Thompson, 2005; Sullivan et al., 2018).

In that regard, Imbens and Pizer (2000) claim that when a key covariate is partly missing, assuming ignorability is inconsistent with random assignment. They instead recommend the use of a constrained, non-ignorable selection model and show that such a model always complies with randomization and encompasses the class of all MAR selection models. Ding and Geng (2014) discuss estimation of subgroup effects when the subgroup-defining covariate has missing values. They consider several selection models - each based on a plausible assumption about the missingness mechanism - and choose the one that best fits the observed data, as evidenced by likelihood ratio tests. However, these works do not present extensions to situations with multiple missing covariates. Horowitz and Manski (2000) and Scharfstein et al. (2011) study the estimation of causal effects in the presence of a missing covariate, without applying identifying assumptions. Their methods focus on obtaining non-parametric bounds for treatment effects, but point identification may often be desirable.

In this thesis, we address situations when multiple baseline variables may be non-ignorably missing. To resolve non-identifiability, we first model the observed data distribution and then non-parametrically identify the full data density using identifying assumptions. A number of identifying assumptions for MNAR settings have been proposed, but we limit our attention to the special *nonparametric saturated (NPS)* (Robins, 1997) family, whose most well-known member is the missing at random restriction. We further draw parallels between imputing missing baseline data and adhering to the protocol of a randomized trial discussed in Rubin (2008). Our aim is to present and contrast multiple imputation performed in the design and the outcome stages of a randomized experiment, respecting and not respecting randomization in each. Throughout, we assume that the treatment and outcome variables are fully observed.

The rest of this thesis is organized as follows. Chapter 2 introduces randomized experiments, associated assumptions as well as observed data modeling for MNAR scenarios. Chapter 3 presents different strategies for imputing missing baseline covariates. Chapter 4 describes the simulation design and results. Chapter 5 concludes with a discussion.

2

Randomized Experiments and Missing Data

2.1 Randomized Experiments

2.1.1 Protocol

Following Rubin (2008), we regard a randomized experiment to have two distinct, successive time points, t_1 and t_2 where $t_1 < t_2$. At t_1, we observe a unit, which may or may not have been exposed to a fixed set of experimental conditions, and its covariates that have been measured before exposure. For simplicity, we consider a two-group design, with an active treatment and a placebo or control. There are two potential outcomes linked to each unit, depending on whether it receives the treatment or the control at t_1. We observe only one of these at t_2, which we call the observed outcome. The times t_1 and t_2 can vary across units, but the length of the interval [t_1,t_2] remains constant. We call t_1 the "design stage" and t_2 the "outcome stage" of the randomized trial. The causal effect of the treatment is any quantity that compares potential outcomes under treatment and control across all units or for a certain subset of units.

2.1.2 *Notation & Assumptions*

Consider a randomized experiment with n subjects. Let T_i be the treatment indicator, where $T_i = 1$ if subject i receives the treatment at t_1 and $T_i = 0$ if subject i receives the control at t_1. Let X_{ij} denote the value of covariate j measured for unit i at baseline, and let $X_j = (X_{1j}, ..., X_{nj})'$ represent measurements for all n subjects. If p covariates are measured, let $X = (X_1, ..., X_p)$. We denote the potential outcomes associated with the treatment and the control by $Y_i(1)$ and $Y_i(0)$ respectively. At t_2, we observe the outcome Y_i. For notational clarity, we forgo the subscript i and use $(T, X, Y(1), Y(0), Y)$ in place of $(T_i, X_i, Y_i(0), Y_i(1), Y_i)$. We work under the following two assumptions.

A1. *Stable Unit Treatment Value Assumption (SUTVA)* (Rubin, 1978).

There is only one version of the treatment and no interference between units. Thus, Y is a deterministic function of $Y(1)$, $Y(0)$ and T, i.e., $Y = TY(1) + (1 - T)Y(0)$.

A2. *Randomization*.

The treatment status T is completely independent of the potential outcomes $Y(0), Y(1)$ as well as the covariates X, which we write as $T \perp\!\!\!\perp \{X, Y(0), Y(1)\}$.

2.2 Modeling MNAR Data

In this section, we momentarily ignore the treatment and the outcome and illustrate observed data modeling for generic variables $X_1, ..., X_p$. Considerations specific to randomized trials are discussed in Chapter 3.

Suppose that $X_1, ..., X_p$ are partly missing. Let $D_1, ..., D_p$ denote their missingness indicators, such that $D_j = 1$ when X_j is missing and $D_j = 0$ when X_j is observed, for $j = 1, ..., p$. Denote $X = (X_1, ..., X_p)$ and $D = (D_1, ..., D_p)'$. Each D represents a missingness pattern, according to which X can be divided into two components. We call these X_{obs}, the observed part of X, and X_{mis}, the missing part of X. If the variables are non-ignorably missing, we need to specify the joint distribution of X and D by making some untestable assumptions about the missingness process. To do this, we use observed data modeling (Linero and Daniels, 2018), where the full data distribution $f(X, D)$ is split into two components - (1) the observed data distribution, $f(X_{obs}, D)$ and (2) the extrapolation distribution $f(X_{mis}|X_{obs}, D)$ (Daniels and Hogan, 2008). $f(X_{obs}, D)$ is identifiable from the observed data, whereas $f(X_{mis}|X_{obs}, D)$ is not. For each missingness pattern D, we specify a model for $f(X_{obs}, D)$. Following this, we can recover $f(X_{mis}, X_{obs}, D)$ from $f(X_{obs}, D)$ by imposing identifying restrictions on $f(X_{mis}|X_{obs}, D)$.

An identifying restriction is an assumption that links $f(X_{obs}, D)$ to $f(X_{mis}|X_{obs}, D)$. As discussed in Linero and Daniels (2018), identifying restrictions can be marginal or joint. A marginal restriction does not identify the full data distribution, but non-parametrically identifies the densities $f(X_j)$, for $j = 1, ..., p$. Examples of marginal restrictions include sequential explainability (Vansteelandt et al., 2007) and nearest identified pattern (Linero, 2017) for non-monotone missingness and last-occasion-pattern-mixture (Birmingham et al., 2003) for monotone missingness. On the other hand, joint restrictions identify the full data density. Such restrictions include the complete case missing value restriction (Little, 1993), the itemwise conditionally independent nonresponse assumption (ICIN) (Sadinle and Reiter, 2017), the pairwise MAR assumption (Tchetgen Tchetgen et al., 2016), the group permutation missingness mechanism of Robins (1997) and also the commonly used MAR assumption (Gill et al., 1997).

A full data distribution identified under a joint restriction is said to be non-parametric saturated (NPS) (Robins, 1997) or non-parametric identified (Vansteelandt et al., 2006), when it meets the following definition.

Definition 2.1 (Non-parametric Saturation)**.** Let $\mathcal{G}$ denote the class of observed data distributions, and let $\mathcal{F}_{\mathcal{A}}$ denote a family of full data distributions under restriction(s) $\mathcal{A}$. $\mathcal{F}_{\mathcal{A}}$ is said to be non-parametric saturated if given $g \in \mathcal{G}$, there exists a unique $f \in \mathcal{F}_{\mathcal{A}}$ that marginalizes to g.

NPS is a favorable property for sensitivity analysis in that each joint restriction strictly relates to the unidentifiable extrapolation distribution, and necessarily preserves the observed data distribution. Hence, if we identify several full data distributions using assumptions in the NPS class, then each of these will exactly agree in their fit to the observed data. Any differences in the results can be attributed to the differing identifying assumptions. For examples of sensitivity analysis using identifying restrictions, see Thijs et al. (2002).

To fix ideas, we consider two binary variables, X_1 and X_2, and show identification under the ICIN and MAR assumptions that lead to non-parametric saturation.

2.2.1 Observed Data Modeling

The vector $D = (D_1, D_2)'$ encodes four missingness patterns, corresponding to which the observed data offer the following nine probabilities

$$
\begin{aligned}
\theta_1 &= P(X_1 = 0, X_2 = 0, D_1 = 0, D_2 = 0),\\
\theta_2 &= P(X_1 = 0, X_2 = 1, D_1 = 0, D_2 = 0),\\
\theta_3 &= P(X_1 = 1, X_2 = 0, D_1 = 0, D_2 = 0),\\
\theta_4 &= P(X_1 = 1, X_2 = 1, D_1 = 0, D_2 = 0),\\
\theta_5 &= P(X_1 = 1, D_1 = 0, D_2 = 1),\\
\theta_6 &= P(X_1 = 0, D_1 = 0, D_2 = 1),\\
\theta_7 &= P(X_2 = 1, D_1 = 1, D_2 = 0),\\
\theta_8 &= P(X_2 = 0, D_1 = 1, D_2 = 0),\\
\theta_9 &= P(D_1 = 1, D_2 = 1).
\end{aligned}
$$

The distribution of $\theta = (\theta_1, ..., \theta_9)'$ corresponds to the observed data density. The full data distribution $f(X_1, X_2, D_1, D_2)$ becomes identifiable by imposing seven independent constraints.

2.2.2 *Identification under ICIN*

Definition 2.2 (Itemwise Conditionally Independent Nonresponse (ICIN))**.** X_1 and X_2 are missing according to the itemwise conditionally independent nonresponse assumption when $X_j \perp\!\!\!\perp D_j | X_{-j}, D_{-j}$ for $j = 1, 2$.

In other words, we surmise that controlling for all other variables and their missingness indicators, the missingness of X_j does not predict its value and vice versa. To identify the full data distribution under ICIN, we follow Theorem 1 from Sadinle and Reiter (2017). Define

$$\eta_D(X_{obs}) = \log f(X_{obs}, D) - \log \int_{X_{mis}} \exp \sum_{D^{*}<D} \eta_{D^*}(X_{obs}) \mathbb{1}(D^* < D) \, dX_{mis}.$$

Here, $D^* < D$ if D has strictly greater number of missing elements than D^*. For each missingness pattern, the η are obtained as follows.

$$\begin{aligned}
\eta_{00}(X_1, X_2) &= \log f(X_1, X_2, D_1 = 0, D_2 = 0), \\
\eta_{01}(X_1) &= \log f(X_1, D_1 = 0, D_2 = 1) - \log \int_{X_2} f(X_1, X_2, D_1 = 0, D_2 = 0) \, dX_2 \\
&= \log \frac{f(X_1, D_1 = 0, D_2 = 1)}{f(X_1, D_1 = 0, D_2 = 0)}, \\
\eta_{10}(X_2) &= \log f(X_2, D_1 = 1, D_2 = 0) - \log \int_{X_1} f(X_1, X_2, D_1 = 0, D_2 = 0) \, dX_1 \\
&= \log \frac{f(X_2, D_1 = 1, D_2 = 0)}{f(X_2, D_1 = 0, D_2 = 0)},
\end{aligned}$$

$$
\begin{aligned}
\eta_{11} &= \log f(D_1 = 1, D_2 = 1) \\
&\qquad - \log \int_{X_1} \int_{X_2} \exp \Big[\eta_{00}(X_1, X_2) + \eta_{01}(X_1) + \eta_{10}(X_2) \Big] dX_2 dX_1 \\
&= \log \frac{f(D_1 = 1, D_2 = 1)}{C},
\end{aligned}
$$

where,

$$
\begin{aligned}
C = \int_{X_1} \int_{X_2} \bigg(& \frac{f(X_1, X_2, D_1 = 0, D_2 = 0) f(X_1, D_1 = 0, D_2 = 1)}{f(X_2, D_1 = 1, D_2 = 0)} \\
& \times \frac{f(X_1, D_1 = 0, D_2 = 0)}{f(X_2, D_1 = 0, D_2 = 0)} \bigg) \, dX_2 dX_1.
\end{aligned}
$$

The full data distribution for each missingness pattern D is given as

$$
f(X, D) = \exp \sum_{D* \leqslant D} \eta_{D*}(X_{obs}).
$$

Hence,

$$
\begin{aligned}
f(X_1, X_2, D_1 = 0, D_2 = 1) &= \exp \Big[\eta_{00}(X_1, X_2) + \eta_{01}(X_1) \Big] \\
&= \frac{f(X_1, X_2, D_1 = 0, D_2 = 0) f(X_1, D_1 = 0, D_2 = 1)}{f(X_1, D_1 = 0, D_2 = 0)}, \\
f(X_1, X_2, D_1 = 1, D_2 = 0) &= \exp \Big[\eta_{00}(X_1, X_2) + \eta_{10}(X_2) \Big] \\
&= \frac{f(X_1, X_2, D_1 = 0, D_2 = 0) f(X_2, D_1 = 1, D_2 = 0)}{f(X_2, D_1 = 0, D_2 = 0)},
\end{aligned}
$$

$$
\begin{aligned}
f(X_1, X_2, D_1 = 1, D_2 = 1) &= \exp\left[\eta_{00}(X_1, X_2) + \eta_{01}(X_1) + \eta_{10}(X_2) + \eta_{11}\right] \\
&= \frac{f(X_1, X_2, D_1 = 0, D_2 = 0) f(X_1, D_1 = 0, D_2 = 1)}{f(X_1, D_1 = 0, D_2 = 0)} \\
&\times \frac{f(X_2, D_1 = 1, D_2 = 0) f(D_1 = 1, D_2 = 1)}{f(X_2, D_1 = 0, D_2 = 0)} \times \frac{1}{C}.
\end{aligned}
$$

2.2.3 Identification under MAR

Definition 2.3 (Missing at Random (MAR))**.** If X_1 and X_2 are Missing at Random, then for each missingness pattern D, $f(D|X_{mis}, X_{obs}) = f(D|X_{obs})$ which can be re-expressed as $f(X_{mis}|X_{obs}, D) = f(X_{mis}|X_{obs})$.

Thus, the extrapolation distribution under a missingness pattern is the one obtained upon marginalizing across all patterns that exist in the data. To obtain an NPS full data distribution, we define $\mathcal{D}$ to be the set of all missingness patterns of (X_1, X_2) that have been realized. For illustration purposes, we assume that $\mathcal{D} = \{(0,0),(0,1),(1,0),(1,1)\}$, i.e., there are a reasonable number of observations for each possible missingness pattern. For each $D \in \mathcal{D}$, we have

$$
f(X_{mis}|X_{obs}, D) = f(X_{mis}|X_{obs}) = \sum_{D^* \in \mathcal{D}} f(X_{mis}|X_{obs}, D^*) f(D^*|X_{obs}).
$$

For $(D_1, D_2) = (0, 1)$, we have

$$
\begin{aligned}
f(X_2|X_1, 0, 1) &= f(X_2|X_1, 0, 0) f(0, 0|X_1) + f(X_2|X_1, 0, 1) f(0, 1|X_1) \\
&+ f(X_2|X_1, 1, 0) f(1, 0|X_1) + f(X_2|X_1, 1, 1) f(1, 1|X_1).
\end{aligned}
$$

For each pattern in the RHS of the above equation, we equate the appropriate extrapolation distribution to the corresponding distribution under $D = (0, 0)$. That is, for $D^* \in \mathcal{D}$, we equate $f(X_{mis}|X_{obs}, D^*) = f(X_{mis}|X_{obs}, 0, 0)$.

Thus, we have

$$
\begin{aligned}
&f(X_2|X_1,0,1) = f(X_2|X_1,0,0),\\
&f(X_2|X_1,1,0) \propto f(X_1,X_2,1,0)\\
&\qquad \propto f(X_1|X_2,1,0)f(X_2,1,0) = f(X_1|X_2,0,0)f(X_2,1,0),\\
&f(X_2|X_1,1,1) \propto f(X_1,X_2,1,1)\\
&\qquad \propto f(X_1,X_2|1,1)f(1,1) = f(X_1,X_2|0,0)f(1,1).
\end{aligned}
$$

We similarly identify the extrapolation distributions for the remaining missingness patterns as shown below. For $(D_1, D_2) = (1, 0)$, we have

$$
\begin{aligned}
f(X_1|X_2,1,0) &= f(X_1|X_2,0,0)f(0,0|X_2) + f(X_1|X_2,0,1)f(0,1|X_2)\\
&\quad + f(X_1|X_2,1,0)f(1,0|X_2) + f(X_1|X_2,1,1)f(1,1|X_2),
\end{aligned}
$$

where,

$$
\begin{aligned}
&f(X_1|X_2,1,0) = f(X_1|X_2,0,0),\\
&f(X_1|X_2,0,1) \propto f(X_1,X_2,0,1)\\
&\qquad \propto f(X_2|X_1,0,1)f(X_1,0,1) = f(X_2|X_1,0,0)f(X_1,0,1),\\
&f(X_1|X_2,1,1) \propto f(X_1,X_2,1,1)\\
&\qquad \propto f(X_1,X_2|1,1)f(1,1) = f(X_1,X_2|0,0)f(1,1).
\end{aligned}
$$

For $(D_1, D_2) = (1, 1)$, we have

$$
\begin{aligned}
f(X_1,X_2|1,1) &= f(X_1,X_2|0,0)f(0,0) + f(X_1,X_2|0,1)f(0,1)\\
&\quad + f(X_1,X_2|1,0)f(1,0) + f(X_1,X_2|1,1)f(1,1),
\end{aligned}
$$

where,

$$
\begin{aligned}
&f(X_1,X_2|1,1) = f(X_1,X_2|0,0),\\
&f(X_1,X_2|0,1) \propto f(X_1,X_2,0,1)\\
&\qquad \propto f(X_2|X_1,0,1)f(X_1,0,1) = f(X_2|X_1,0,0)f(X_1,0,1),\\
&f(X_1,X_2|1,0) \propto f(X_1,X_2,1,0)\\
&\qquad \propto f(X_1|X_2,1,0)f(1,1) = f(X_1|X_2,0,0)f(X_2,1,0).
\end{aligned}
$$

2.2.4 *Multiple Imputation (MI)*

In multiple imputation, missing values are filled in $M > 1$ times, by drawing from the posterior predictive distribution of the missing data given the observed data. In MNAR settings, this is equivalent to independent draws from the extrapolation distribution $f(X_{mis}|X_{obs}, D)$, according to the missingness pattern D. This creates M complete datasets, each of which is analyzed separately. Results are then combined using the rules provided in Rubin (1987), which we now summarize.

Suppose Q is the quantity that we want to estimate. Let q be an estimator of Q, and let u be the estimator of the variance associated with q. From each completed dataset, we obtain q_m and u_m, where $m = 1, .., M$. The point estimate of Q is $\bar{Q}_M = \frac{\sum_{m=1}^{M} q_m}{M}$. The combined within-imputation variance is given by $\bar{U}_M = \frac{\sum_{m=1}^{M} u_m}{M}$ and the between-imputation variance is $B_m = \frac{\sum_{m=1}^{M}(q_m - \bar{Q}_M)^2}{M-1}$. The variance associated with $\bar{Q}_M$ is estimated by $T_M = (1 + \frac{1}{M})B_M + \bar{U}_M$. A $100(1-\alpha)\%$ confidence interval for Q based on the t distribution is $\bar{Q}_M \pm t_{\nu_M, \frac{\alpha}{2}}\sqrt{T_M}$, where $\nu_M = (1 + \frac{1}{r_M})^2(M-1)$ are the degrees of freedom and $r_M = (1 + \frac{1}{M})\frac{B_M}{\bar{U}_M}$.

3

Imputing Missing Baseline Covariates

3.1 Design Stage Imputation

An important attribute of randomized experiments is that they are prospective in nature, i.e., designed *a priori*, without any outcome data in sight (Rubin, 2007). As has been highlighted in Rubin (2002, 2007, 2008), this nature is of significant consequence for ensuring fair causal comparisons. Covariates X are design stage quantities, and should ideally remain free of any influence from the outcome. If we adhere to this, then the imputation model for X will include quantities observed at time t_1 only. In our example, the full data vector will thus be (X_1, X_2, D_1, D_2, T). This forms a 2^5 contingency table.

Covariates are also typically measured at baseline, before randomizing units to treatment or control. Thus, arguably, in most scenarios it is unlikely that their missingness depends on treatment group membership. Because of randomization, it is often reasonable to assume that the distribution of missing values is also balanced across the treatment and control groups. The missingness indicators D_1 and D_2 can therefore be considered "pretreatment" variables, i.e., randomization implies that $(X_1, X_2, D_1, D_2) \perp\!\!\!\perp T$. Owing to this independence, we can collapse the 2^5 table for (X_1, X_2, D_1, D_2, T) across the treatment indicator T. Thus, imputing missing baseline data respecting randomization is equivalent to identifying the full data distribution $f(X_1, X_2, D_1, D_2)$, and drawing imputations from $f(X_{mis}|X_{obs}, D_1, D_2)$. The ICIN and MAR assumptions used to identify $f(X_1, X_2, D_1, D_2)$ will be $X_j \perp\!\!\!\perp D_j|X_{-j}, D_{-j}$ and $f(X_{mis}|X_{obs}, D_1, D_2) = f(X_{mis}|X_{obs})$, respectively.

In view of preserving arm-specific proportions, some works advocate imputing missing covariates within treatment arms (e.g., see Schemper & Smith, 1990). White and Thompson (2005), however, claim the contrary, and use theoretical arguments to show that mean imputation within arms can be unbiased, but slightly inefficient. In the context of our example, not respecting randomization means disregarding the independence of X and D with respect to T, which is equivalent to retaining the 2^5 contingency table and obtaining the full data distribution $f(X_1, X_2, D_1, D_2, T)$. Imputations are then generated from $f(X_{mis}|X_{obs}, D_1, D_2, T)$, analogous to separately imputing within $T = 0$ and $T = 1$. The ICIN and MAR assumptions at play will also condition on the binary treatment indicator, i.e., ICIN will imply $X_j \perp\!\!\!\perp D_j|X_{-j}, D_{-j}, T$ and MAR will imply $f(X_{mis}|X_{obs}, D_1, D_2, T) = f(X_{mis}|X_{obs}, T)$.

If we briefly consider the extrapolation distributions under the methods discussed above, we see that $f(X_{mis}|X_{obs}, D_1, D_2) \propto f(X_{mis}, X_{obs}, D_1, D_2)$ and $f(X_{mis}|X_{obs}, D_1, D_2, T)$ $\propto f(X_{mis}, X_{obs}, D_1, D_2|T)$. By virtue of randomization, the distribution of the observed pretreatment variables will be asymptotically balanced, i.e., as n increases, $f(X_{obs}, D_1, D_2)$ and $f(X_{obs}, D_1, D_2|T)$ become approximately equal. In turn, so do $f(X_{mis}, X_{obs}, D_1, D_2)$ and $f(X_{mis}, X_{obs}, D_1, D_2|T)$. Although chance imbalances may still exist, they will be trivial if the randomization procedure has worked well. Hence, for a moderate to large n, we can expect similar imputation distributions under the methods respecting and not respecting randomization. Nonetheless, this is a conjecture that warrants investigation, particularly for subgroup treatment effect estimation. That being said, if the trial size is very small, then severe imbalance in the observed baseline data may exist. In such a situation, it is possible that imputing within treatment arms worsens overall balance and leads to biased estimates.

3.2 Outcome Stage Imputation

For missing covariates in regression models, the recommended approach is to impute using all other covariates and the outcome (Rubin & Schencker, 1991; Vach & Blettner, 1991; Vach, 1994; Greenland & Finkle, 1995; Schafer, 1997; Barnard & Meng, 1999; Little & Rubin, 2002; Allison, 2002; Moons, et al., 2006). In randomized trials, this has the disadvantage of allowing the outcome to directly influence the design stage, which can bring the face validity of the final conclusions into question. On the other hand, not controlling for Y in imputations can lead to distorted estimates. If a partly missing covariate X is highly predictive of Y, then Y will carry information about X that may not be captured by other variables in the imputation model. If X is imputed without using Y, then the imputed part of X will have no (conditional) association with Y. This could falsely attenuate the overall covariate-outcome association. As discussed in Little (1992), when imputations are drawn from the conditional distribution of X_{mis} given only the observed covariates X_{obs}, the partial correlation between Y and X_{mis} given X_{obs} tends to be ignored, which can result in highly biased estimates.

In the framework of a randomized trial, imputation in the outcome stage will involve a full data vector of quantities observed at times t_1 as well as t_2, i.e., $(X_1, X_2, D_1, D_2, T, Y)$, which forms a contingency table with 2^6 cells. One way to allow randomization to play a role here is to collapse this 2^6 table across T. However, this makes a strong assumption that $Y \perp\!\!\!\perp T$, which may ultimately underestimate the treatment effect. A more principled way would be to factorize the extrapolation distribution $f(X_{mis}|X_{obs}, D_1, D_2, T, Y)$ to naturally represent the design and outcome stages of a randomized experiment. We have

$$\begin{aligned} f(X_{mis}|X_{obs}, D_1, D_2, T, Y) \propto & f(X_{mis}, X_{obs}, D_1, D_2, T, Y) \\ \propto & f(X_{mis}, X_{obs}, D_1, D_2, T)\ f(Y|X_{mis}, X_{obs}, D_1, D_2, T). \end{aligned}$$

Here, $f(X_{mis}, X_{obs}, D_1, D_2, T)$ represents the joint likelihood of the design stage quantities and $f(Y|X_{mis}, X_{obs}, D_1, D_2, T)$ represents the outcome model. Randomization implies that $(X_1, X_2, D_1, D_2) \perp\!\!\!\perp T$, and hence,

$$f(X_{mis}, X_{obs}, D_1, D_2, T) = f(X_{mis}, X_{obs}, D_1, D_2)f(T).$$

Thus, we can impute using

$$f(X_{mis}|X_{obs}, D_1, D_2, T, Y) \propto f(X_{mis}, X_{obs}, D_1, D_2)\ f(Y|X_{mis}, X_{obs}, D_1, D_2, T).$$

To implement this method, we first identify the full data density $f(X_{mis}, X_{obs}, D_1, D_2)$ under an identifying restriction of choice. The extrapolation distribution is then calculated by weighting this density by the conditional distribution of Y, as obtained from the estimated outcome model. We note that this method can be sensitive to the form of the outcome equation specified. Typically, analysis models of interest do not explicitly define Y as a function of the missingness indicators D, in which case, the conditional independence assumption $Y \perp\!\!\!\perp D|X, T$ is implicitly made. In this case, Y is indirectly related to D, via its association with X. If Y is directly related to the missingness in X, i.e., if the D are predictive of Y, then it may be preferable to use this method with an outcome model that includes D.

If we ignore randomization, we essentially use both the treatment and the outcome in the imputation model. This amounts to identifying $f(X_1, X_2, D_1, D_2, T, Y)$ and imputing from $f(X_{mis}|X_{obs}, D_1, D_2, T, Y)$. The ICIN and MAR assumptions will now be $X_j \perp\!\!\!\perp D_j|X_{-j}, D_{-j}, T, Y$ and $f(X_{mis}|X_{obs}, D_1, D_2, T, Y) = f(X_{mis}|X_{obs}, T, Y)$, respectively. Although this method takes no account of the two-stage protocol discussed in Rubin (2008), it is commonly used for filling in missing covariates. This could be because often, very little is known about the true missingness mechanism. Inclusion of as many variables as possible in the imputation model is thus advised (Rubin, 1996).

4

Simulation Study

In this chapter, we conduct repeated sampling studies to evaluate the performance of the four imputation methods presented in Chapter 3. We use twelve simulation settings, comprising all combinations of three missingness scenarios, two identifying assumptions and two covariate-outcome associations. We begin by describing the data generation process.

4.1 Data Generation

We mimic a randomized experiment with $n = 1000$ units. Every unit is associated with two binary covariates, (X_1, X_2). Each unit is assigned to the treatment or the control arm with equal probabilities, i.e., T is generated from a $Ber(0.5)$ distribution. (X_1, X_2) and (D_1, D_2) are generated as per the ICIN and MAR assumptions, under three missingness scenarios.

In the first scenario, we assume that the measurement of covariates and any missingness in them occurs before treatment assignment. Both (X_1, X_2) and (D_1, D_2) are regarded as pretreatment variables and generated completely independent of T. Such a scenario could arise in any well-designed randomized trial. We call this S_{Ind}.

To simulate a missing at random situation under S_{Ind}, we first generate $X_1 \sim Ber(0.8)$ and $X_2 \sim Ber(0.6)$. Subsequently, we delete (at random) 35 $\%$ of the observations from X_1 and 40 $\%$ of the observations from X_2. The true missingness mechanism is therefore MCAR, which is a special case of MAR.

To generate missingness as per ICIN, we make use of the relationship between the ICIN assumption and hierarchical loglinear models, which we first illustrate. The 2^4 table for (X_1, X_2, D_1, D_2) can be split into four, partial 2^2 contingency tables for (X_1, X_2), controlling for the missingness pattern (D_1, D_2). Let $m_{x_1,x_2|(d_1,d_2)}$ represent the count corresponding to $(X_1 = x_1, X_2 = x_2)$ in the partial table with $(D_1, D_2) = (d_1, d_2)$. The conditional independence assumptions represented by ICIN can be expressed in the form of a loglinear model for each $m_{x_1,x_2|(d_1,d_2)}$, as shown below. Mathematical justification for this is presented in Appendix A.

$$\log m_{x_1,x_2|(0,0)} = \lambda_0 + \lambda_{X_1}x_1 + \lambda_{X_2}x_2 + \lambda_{X_1X_2}x_1x_2, \tag{1}$$

$$\log m_{x_1,x_2|(0,1)} = \lambda_0 + \lambda_{X_1}x_1 + \lambda_{X_2}x_2 + \lambda_{D_2}d_2 + \lambda_{X_1X_2}x_1x_2 + \lambda_{X_1D_2}x_1d_2, \tag{2}$$

$$\log m_{x_1,x_2|(1,0)} = \lambda_0 + \lambda_{X_1}x_1 + \lambda_{X_2}x_2 + \lambda_{D_1}d_1 + \lambda_{X_1X_2}x_1x_2 + \lambda_{X_2D_1}x_2d_1, \tag{3}$$

$$\begin{aligned}\log m_{x_1,x_2|(1,1)} = {} & \lambda_0 + \lambda_{X_1}x_1 + \lambda_{X_2}x_2 + \lambda_{D_1}d_1 + \lambda_{D_2}d_2 + \lambda_{X_1X_2}x_1x_2 \\ & + \lambda_{X_1D_2}x_1d_2 + \lambda_{X_2D_1}x_2d_1 + \lambda_{D_1D_2}d_1d_2.\end{aligned} \tag{4}$$

For a given pattern, multinomial probabilities for all combinations of $(X_1 = x_1, X_2 = x_2)$ can be obtained by exponentiating (Agresti, 2012), i.e.,

$$\pi_{x_1,x_2|(d_1,d_2)} = \frac{m_{x_1,x_2|(d_1,d_2)}}{\sum_{x_1}\sum_{x_2} m_{x_1,x_2|(d_1,d_2)}}. \tag{5}$$

We use (1) - (5) for generating data under ICIN as follows. We first draw $D_1 \sim Ber(0.35)$ and $D_2 \sim Ber(0.40)$. The coefficients λ are set as $\lambda_0 = 0.1$, $\lambda_{X_1} = 1.5$, $\lambda_{X_2} = 0.02$, $\lambda_{D_1} = 0.009$, $\lambda_{D_2} = 0.5$, $\lambda_{X_1D_2} = 0.075$, $\lambda_{X_2D_1} = 1$, $\lambda_{X_1X_2} = 0.06$ and $\lambda_{D_1D_2} = 0.25$. Following this, we obtain $m_{x_1,x_2|(d_1,d_2)}$, $\forall(x_1, x_2) \in \{0, 1\}^2$ and

Table 4.1: Counts for the observed data when (X_1, X_2) and (D_1, D_2) are generated as per ICIN, using $\lambda_0 = 0.1$, $\lambda_{X_1} = 1.5$, $\lambda_{X_2} = 0.02$, $\lambda_{D_1} = 0.009$, $\lambda_{D_2} = 0.5$, $\lambda_{X_1 D_2} = 0.075$, $\lambda_{X_2 D_1} = 1$, $\lambda_{X_1 X_2} = 0.06$ and $\lambda_{D_1 D_2} = 0.25$. Data is generated for one particular random seed. NA represents missing values.

	$\boldsymbol{X_2 = 0}$	$\boldsymbol{X_2 = 1}$	$\boldsymbol{X_2 =}$ **NA**
$\boldsymbol{X_1 = 0}$	40	35	48
$\boldsymbol{X_1 = 1}$	175	169	208
$\boldsymbol{X_1 =}$ **NA**	49	154	122

$\forall (d_1, d_2) \in \{0, 1\}^2$, using the appropriate equation from (1) - (4). For each pattern $(D_1 = d_1, D_2 = d_2)$, we jointly draw $(X_1, X_2) = (x_1, x_2)$ with probability $\pi_{x_1, x_2 | (d_1, d_2)}$ and set X_j to missing wherever $D_j = 1$, where $j = 1, 2$. Table 4.1 displays the counts for the resultant observed data when using this approach.

In the second scenario, we simulate (X_1, X_2) independent of T, but (D_1, D_2) are generated conditional on the treatment assignment. This can happen, for example, when baseline observations specific to a treatment arm are lost. Covariates could also be variables like gender or age, which are anyway independent of the treatment. If such variables are measured post-randomization, then this situation can arise if one treatment group co-operates better than the other during data collection. We call this scenario S_T. We generate data under S_T using an approach similar to S_{Ind}, except that D_1 and D_2 are now drawn from the conditional distribution $Pr(D_1 = 1 | T = 1) = 0.35$, $Pr(D_1 = 1 | T = 0) = 0.1$, $Pr(D_2 = 1 | T = 1) = 0.40$ and $Pr(D_2 = 1 | T = 0) = 0.25$.

In the third scenario, we generate (X_1, X_2) and (D_1, D_2) completely independent of T, but (D_1, D_2) are now predictive of Y. This scenario could occur in longitudinal trials, with selective loss to follow up (Moons et al., 2006). We call this S_Y. Under S_Y, (X_1, X_2) and (D_1, D_2) are generated in the exact same manner as S_{Ind}. The non-zero (conditional) association between Y and (D_1, D_2) is reflected in the outcome generation model, as explained further.

The outcome Y is generated using

$$g(Pr(Y = 1|X_1, X_2, T)) = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_t T_+ \alpha_{tx_2} T X_2 + \alpha_{D_1} D_1 + \alpha_{D_2} D_2,$$

where $g(.)$ denotes the logit link function. Here, α_t is set to 0.3. α_{D_1} and α_{D_2} are 0 for scenarios S_{Ind} and S_T and -0.6 and -0.4 for S_Y. We vary the strength of the association between (X_1, X_2) and Y and set the other coefficients accordingly.

High Association To reflect a situation where the baseline variables are highly predictive of the outcome, we set $OR(Y, X_1|T, X_2) = 2$, $OR(Y, X_2|T = 0, X_1) = 2.5$ and $OR(Y, X_2|T = 1, X_1) = 4$, which leads to

$$g(Pr(Y = 1|X_1, X_2, T)) = 0.8X_1 + 0.9X_2 + 0.3T + 0.5TX_2.$$

Low Association We set $OR(Y, X_1|T, X_2) = OR(Y, X_2|T = 1, X_1) = OR(Y, X_2|T = 0, X_1) \approx 1$, which gives

$$g(Pr(Y = 1|X_1, X_2, T)) = 0.02X_1 + 0.05X_2 + 0.3T + 0.015TX_2.$$

4.2 Methods

For analysis, we specify a working model that adjusts for X_1 and X_2, and introduce effect modification with respect to X_2. We have

$$\text{logit}\ (Pr(Y = 1|X_1, X_2, T)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_t T + \beta_{tx_2} T X_2.$$

For scenarios S_{Ind} and S_T, this model has the same form as the outcome generation model, except that we permit a non-zero intercept. For S_Y, the analysis model will be misspecified. The adjusted treatment effect (on the log-odds scale), heterogeneous across categories of X_2, will be given by $\beta_t + \beta_{tx_2} x_2$. The parameters of interest are hence β_t and β_{tx_2}.

We create multiply imputed datasets according to the four methods discussed in Chapter 3 - respecting randomization in the design stage (R), not respecting randomization in the design stage (NR), respecting randomization in the outcome stage (RY) and not respecting randomization in the outcome stage (NRY). For method R, there are nine observed probabilities, $\theta = \{\theta_1, ..., \theta_9\}$, from the contingency table for (X_1, X_2, D_1, D_2, T) collapsed across T. We treat the corresponding counts $\{n_1, ..., n_9\}$ as a multinomial sample and place a Dirichlet ($\frac{1}{9}$) prior on θ. For method NR, we have eighteen observed probabilities, $\theta = \{\theta_1, ..., \theta_{18}\}$, with counts $\{n_1, ..., n_{18}\}$, nine in each treatment arm. Here, we use a Dirichlet($\frac{1}{18}$) prior for θ. Similarly, for NRY, θ is a vector of thirty six observed probabilities, $\theta = \{\theta_1, ..., \theta_{36}\}$, nine in each category of (Y, T), on which we place a Dirichlet $(\frac{1}{36})$ prior.

We carry out multiple imputation in two steps. We first sample a value for θ from its posterior distribution, from which we obtain the extrapolation distributions under the chosen identifying restriction. For each missing data point, we then generate an imputation from the pattern-specific extrapolation distribution. We repeat these two steps $M = 100$ independent times, creating 100 completed datasets. For each of these, we obtain point and variance estimates for β_t and β_{tx_2}, and combine them. For method RY, we make Bayesian updates in two blocks. In one block, we sample $\theta = \{\theta_1, ..., \theta_9\}$ from its posterior distribution, and calculate the full data density $f(X_1, X_2, D_1, D_2)$. Using this and the currrent value of the outcome model parameters $\beta = (\beta_0, \beta_1, \beta_2, \beta_t, \beta_{tx_2})'$, we generate an imputed dataset using $f(X_{mis}|X_{obs}, D_1, D_2, T, Y) \propto f(X_{mis}, X_{obs}, D_1, D_2) f(Y|X_{obs}, X_{mis}, D_1, D_2, T)$. In the other block, we use this dataset to update β from its posterior distribution. For sampling β, we use the Polya-Gamma data augmentation technique outlined in Polson et al. (2013).

We repeat all simulations 1000 times and evaluate the performance of each method based on the absolute bias, the multiple imputation standard error (MI-SE), the Monte Carlo standard deviation (MC-SD), and coverage of the 95 $\%$ confidence intervals (95 $\%$ credible intervals for RY).

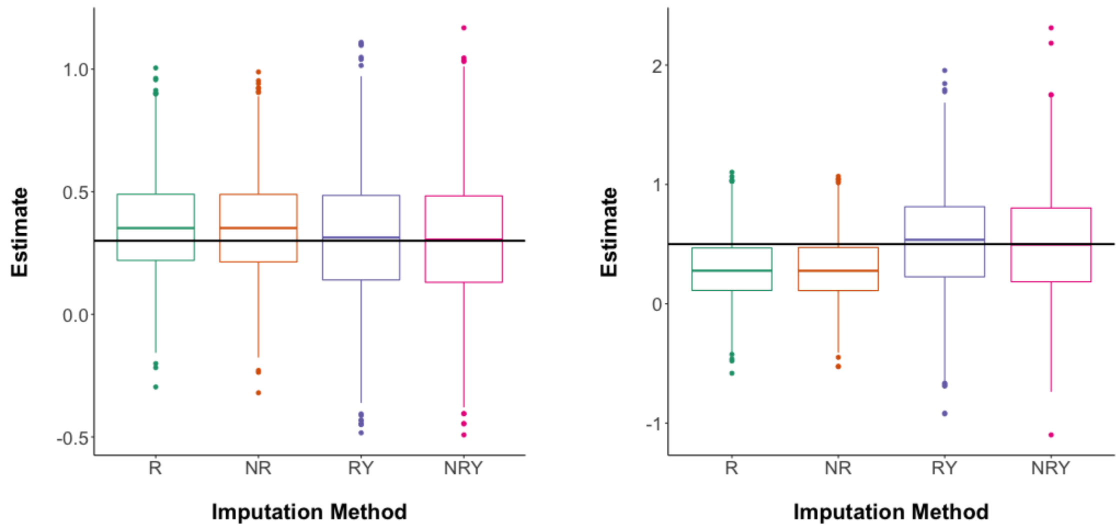

FIGURE 4.1: The left panel presents the distribution of the β_t estimates over 1000 replications, where the true value of $\beta_t = 0.3$. The right panel represents the distribution of the β_{tx_2} estimates over 1000 replications, where the true value of $\beta_{tx_2} = 0.5$. Estimates for both coefficients have been obtained under ICIN for the S_{Ind}, high association setting.

4.3 Results

In this section, we present results under the ICIN assumption. Results under MAR are similar to those under ICIN, and are displayed in Appendix B.

4.3.1 S_{Ind} : Independent Missingness

Figures 4.1 and 4.2 summarize coefficient estimates under S_{Ind}, for the high and low association settings respectively. Tables 4.2 and 4.3 display the standard errors and coverage probabilities for the same.

For both coefficients, R and NR produce biased, but almost equivalent point estimates. We also see little difference in the standard errors between the approaches. We ascribe this finding to the large sample size of 1000 units. As explained in section 3.1, the balance achieved by randomization in this case is near-perfect, due to which methods using and not using T tend to have similar extrapolation distributions and consequently, similar imputations. We also note that the analysis model we use adjusts for both X_1 and X_2.

Adjustment for covariates is known to eliminate bias due to any residual imbalance that might exist in spite of randomization. Hence, trivial imbalances in covariate distributions induced by imputations under NR do not seem to have a sizeable impact on the estimates of β_t or β_{tx_2}. Comparing RY and NRY, we see a similar pattern, although biases under RY are slightly more pronounced than NRY (0.01 versus 0.006 for β_t and 0.02 versus 0.006 for β_{tx_2}). Standard errors under the two methods are approximately equal.

For β_t, we observe that the outcome stage methods produce approximately unbiased estimates (absolute bias $\leqslant$ 0.02) and biases for their design stage counterparts are only slightly higher (≈ 0.05). Differences are much more pronounced for the interaction coefficient β_{tx_2}, with biases for methods R and NR escalating to 0.2 in the negative direction. β_{tx_2} is a measure of the conditional (Y, X_2) association, i.e., $\beta_{tx_2} = OR(Y, X_2|T = 1, X_1)$ which becomes diluted as a result of imputing X_2 without using Y, explaining the large negative bias. The effect of excluding the outcome in imputations is less evident in β_t, which is equivalent to $OR(Y, T|X_1, X_2)$ and reflects the conditional association between Y and T rather than Y and X_2. The design stage methods do exhibit modest efficiency advantages over the outcome stage methods, with empirical standard errors for β_{tx_2} about 25 % smaller.

For the design stage methods, the MI-SEs for both coefficients are fairly large compared to the corresponding Monte Carlo standard errors. Particularly for β_{tx_2}, the standard error is large even relative to the point estimate. Although the MI variance estimator T_M can be positively biased (Wang and Robins, 1998; Robins & Wang, 2000; Reiter & Raghunathan, 2007), in this case, the mean value of T_M is almost 52 % larger than the corresponding Monte Carlo variance. Given the diluted estimates for β_{tx_2}, such large standard errors lead to wide confidence intervals containing 0 (almost 99 % of the times in the simulations). This has implications when testing for the significance of the treatment effect. We tend to make many Type II errors and not find a significant effect where it should be found. Sterne et al. (2009) provide an example of one such randomized trial, where a null association

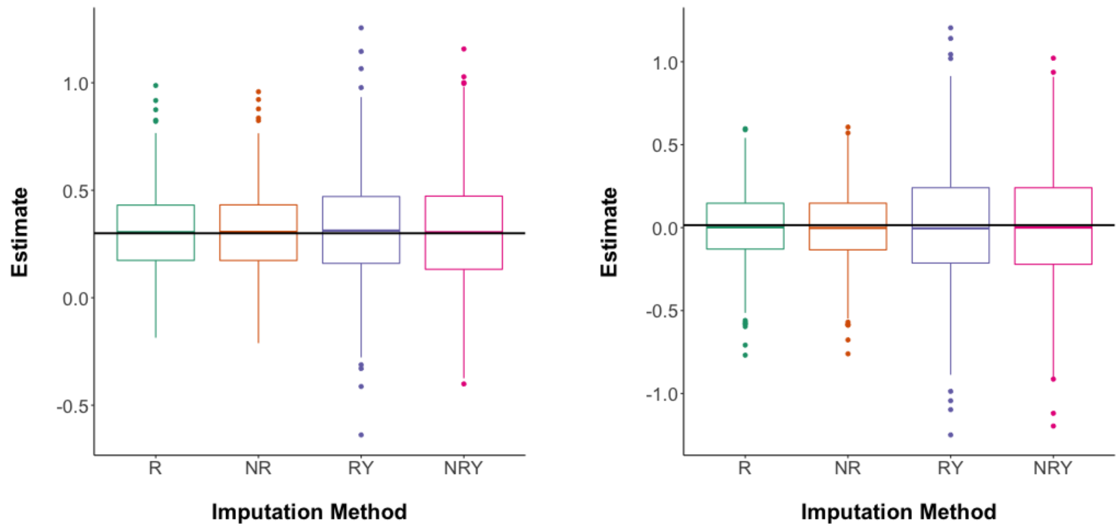

FIGURE 4.2: The left panel represents the distribution of the β_t estimates over 1000 replications, where the true value of $\beta_t = 0.3$. The right panel represents the β_{tx_2} estimates over 1000 replications, where the true value of $\beta_{tx_2} = 0.015$. Estimates for both coefficients have been obtained under ICIN for the S_{Ind}, low association setting.

between the exposure and outcome was found due to the omission of the outcome from the covariate imputation model. In terms of empirical coverage, R and NR seem to overcover, with coverage rates around 0.97-0.98, whereas RY and NRY exhibit close to nominal coverage.

In the low association setting, methods respecting and not respecting randomization once again seem to produce equivalent estimates and standard errors relative to each other. We see that both, the design and the outcome stage methods, produce unbiased coefficient estimates (absolute bias < 0.01 for all). Owing to the low association, the outcome does not carry much information about the partly missing covariates, which reduces the omitted variable bias caused due to its exclusion in the imputation model. The MI-SEs for β_{tx_2} are largely inflated for all the four imputation methods. This is not surprising, since adjustment for non-prognostic "random-noise" covariates in logistic regression leads to a loss in precision when estimating the treatment effect (Robinson & Jewell, 1991; Kahan et al., 2014). Although at the expense of wider confidence intervals (for β_{tx_2}), RY and NRY

Table 4.2: Standard errors and coverage rates for β_t and β_{tx_2} under the S_{Ind}, high association setting.

	β_t			β_{tx_2}		
Method	MC-SD	MI-SE	Coverage	MC-SD	MI-SE	Coverage
R	0.2057	0.2501	0.98	0.2629	0.3846	0.979
NR	0.2056	0.2507	0.981	0.2625	0.3850	0.982
RY	0.2599	0.2515	0.936	0.4476	0.4349	0.942
NRY	0.2587	0.2530	0.943	0.4591	0.4411	0.938

Table 4.3: Standard errors and coverage rates for β_t and β_{tx_2} under the S_{Ind}, low association setting.

	β_t			β_{tx_2}		
Method	MC-SD	MI-SE	Coverage	MC-SD	MI-SE	Coverage
R	0.1812	0.2243	0.986	0.209	0.3077	0.998
NR	0.1815	0.2244	0.985	0.209	0.3078	0.997
RY	0.2364	0.2380	0.953	0.3363	0.3341	0.953
NRY	0.2463	0.2385	0.944	0.3340	0.3361	0.949

exhibit nominal coverage but R and NR still overcover.

4.3.2 S_T : Treatment Dependent Missingness

Figures 4.3 and 4.4 summarize coefficient estimates under S_T for the high and low association settings respectively. Tables 4.4 and 4.5 provide standard errors and coverage probabilities for the estimates.

As before, we see that methods respecting and not respecting randomization produce similar coefficient and standard error estimates. Imputation without the outcome seems to have a much larger impact than in the previous scenario. When the covariate-outcome

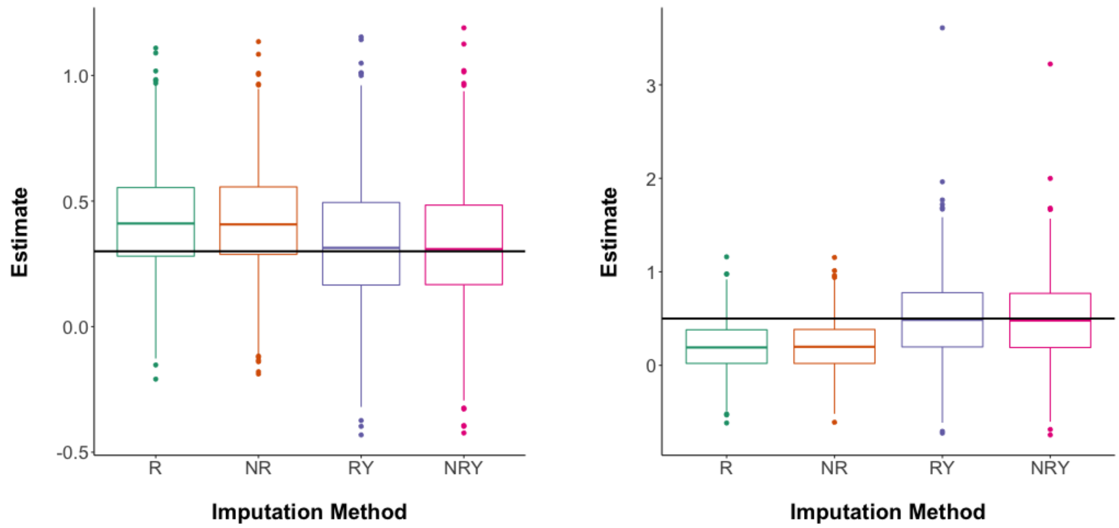

FIGURE 4.3: The left panel represents the distribution of the β_t estimates over 1000 replications, where the true value of $\beta_t = 0.3$. The right panel represents the distribution of the β_{tx_2} estimates over 1000 replications, where the true value of $\beta_{tx_2} = 0.5$. Estimates for both coefficients have been obtained under ICIN for the S_T, high association setting.

association is high, estimates for β_t under R and NR are substantially inflated, with absolute biases around 0.1. Estimates for β_{tx_2} seem under to be much more attenuated towards 0. Even for the low association setting, we see that the design stage methods seem to exhibit larger biases than the outcome stage methods. This is probably because Y affects the missingness in (X_1, X_2) in two ways - via its association with (X_1, X_2) and also via its strong association with T. T in turn, is directly associated with the missingness in the covariates. Standard errors under all four methods behave similarly as under S_{Ind}.

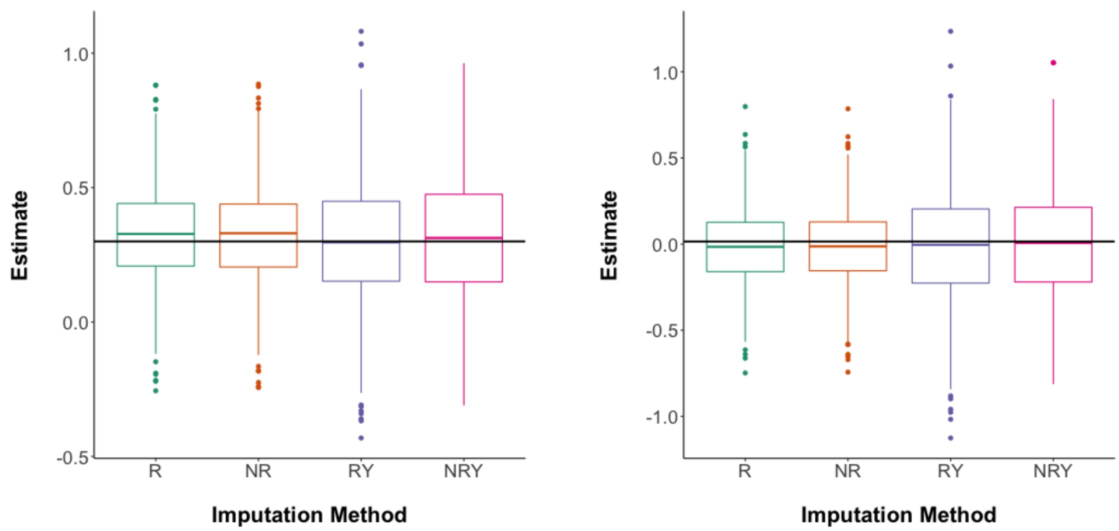

FIGURE 4.4: The left panel represents the distribution of the β_t estimates over 1000 replications, where the true value of $\beta_t = 0.3$. The right panel represents the distribution of the β_{tx_2} estimates over 1000 replications, where the true value of $\beta_{tx_2} = 0.015$. Estimates for both coefficients have been obtained under ICIN for the S_T, high association setting.

Table 4.4: Standard errors and coverage rates for β_t and β_{tx_2} under the S_T, high association setting.

	β_t			β_{tx_2}		
Method	MC-SE	MI-SE	Coverage	MC-SE	MI-SE	Coverage
R	0.2016	0.2436	0.965	0.2649	0.3789	0.951
NR	0.2023	0.2422	0.964	0.2663	0.3795	0.953
RY	0.2426	0.2429	0.946	0.4415	0.4286	0.943
NRY	0.2435	0.2438	0.949	0.4412	0.4281	0.936

Table 4.5: Standard errors and coverage rates for β_t and β_{tx_2} under the S_T, low association setting.

	β_t			β_{tx_2}		
Method	MC-SE	MI-SE	Coverage	MC-SE	MI-SE	Coverage
R	0.1755	0.2183	0.981	0.2088	0.3012	0.993
NR	0.1757	0.2185	0.982	0.2088	0.3015	0.952
RY	0.2255	0.2286	0.942	0.3184	0.3321	0.952
NRY	0.2391	0.2222	0.924	0.3233	0.3171	0.934

4.3.3 S_Y : Outcome Dependent Missingness

As mentioned before, generating Y as a function of D_1 and D_2 leads to a misspecified outcome model for analysis. Here, we follow Sullivan et al.(2018) and base our conclusions on the "least false" values of β_t and β_{tx_2} obtained by fitting the misspecified model to the complete data (abbreviated ‘NoMiss‘ in Figures 4.5 & 4.6 and Tables 4.6 & 4.7). Our observations closely follow those under S_{Ind}, with some minor distinctions. Once again, R and NR produce almost equal point estimates and standard errors. The bias for NRY is almost 0 for both coefficients, but is slightly higher for RY. This may be because results under RY depend on how strongly $Y \perp\!\!\!\perp (D_1, D_2)|X_1, X_2, T$ holds. In this scenario, the outcome is generated a as a function of (D_1, D_2), which violates that assumption. In the outcome generation model, we set $OR(Y, D_1|X_1, X_2, T) = 0.54$ and $OR(Y, D_2|X_1, X_2, T) = 0.67$, but we would expect the observed bias to increase if D_1 and D_2 have a higher conditional association with Y.

For β_t and β_{tx_2}, imputing without Y once again produces biased estimates when the covariate-outcome association is high, although absolute biases are slightly more marked compared to S_{Ind}. We attribute this to the fact that Y now influences the missingness in X in a two-fold manner, by means of D as well as its (high) association with X.

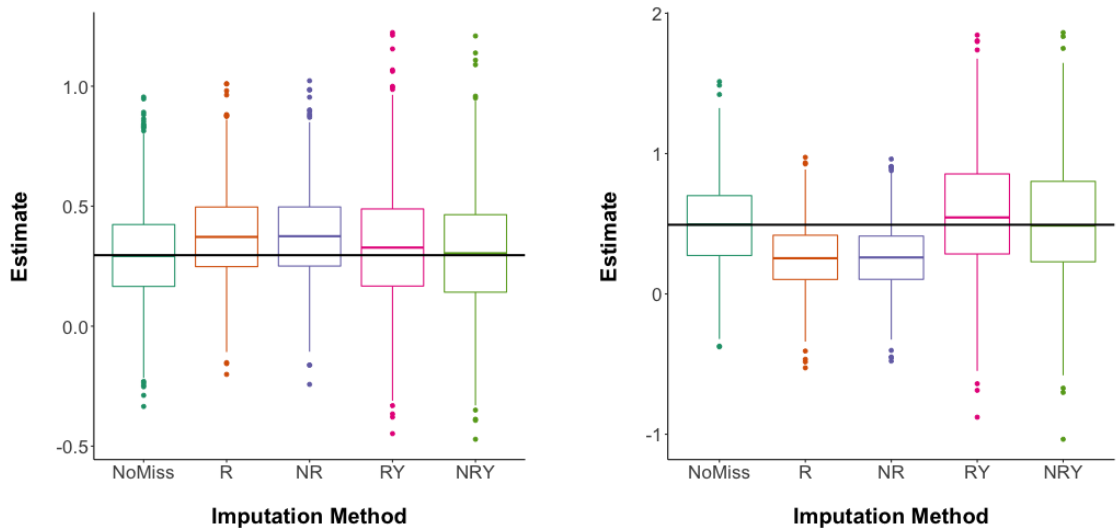

FIGURE 4.5: The left panel represents the distribution of the β_t estimates over 1000 replications, where the true value of $\beta_t = 0.3$. The right panel represents the distribution of the β_{tx_2} estimates over 1000 replications, where the true value of $\beta_{tx_2} = 0.5$. Estimates for both coefficients have been obtained under ICIN for the S_Y, high association setting.

Turning to the low association setting, all methods produce unbiased estimates for both coefficients, with absolute bias for RY slightly larger than the others. Standard errors and coverage rates for all methods are akin to the ones obtained under S_{Ind}.

Table 4.6: Standard errors and coverage rates for β_t and β_{tx_2} under the S_Y, high association setting.

	β_t			β_{tx_2}		
Method	MC-SD	MI-SE	Coverage	MC-SD	MI-SE	Coverage
R	0.1878	0.2392	0.978	0.2299	0.3522	0.978
NR	0.1884	0.2395	0.982	0.2299	0.3527	0.98
RY	0.2457	0.2468	0.951	0.4116	0.4045	0.946
NRY	0.2443	0.2448	0.95	0.4207	0.4116	0.955
NoMiss	0.2036	0.2095	0.95	0.3022	0.3036	0.95

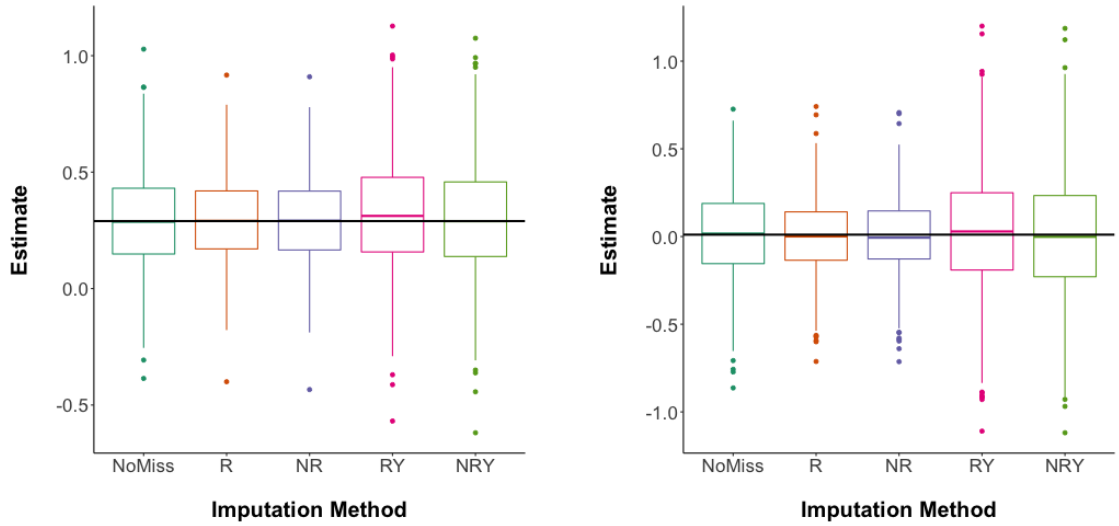

FIGURE 4.6: The left panel represents the distribution of the β_t estimates over 1000 replications, where the true value of $\beta_t = 0.3$. The right panel represents the distribution of the β_{tx_2} estimates over 1000 replications, where the true value of $\beta_{tx_2} = 0.015$. Estimates for both coefficients have been obtained under ICIN for the S_Y, low association setting.

Table 4.7: Standard errors and coverage rates for β_t and β_{tx_2} under the S_Y, low association setting.

	β_t			β_{tx_2}		
Method	MC-SD	MI-SE	Coverage	MC-SD	MI-SE	Coverage
R	0.1783	0.2233	0.989	0.2093	0.3063	0.996
NR	0.1795	0.2040	0.986	0.2115	0.3073	0.994
RY	0.2412	0.2362	0.942	0.3390	0.3311	0.948
NRY	0.2412	0.2382	0.949	0.343	0.3356	0.942
NoMiss	0.2029	0.2019	0.949	0.2599	0.2460	0.95

5

Discussion

In this thesis, we compare four different ways of multiply imputing missing covariates in randomized experiments. We conduct repeated sampling evaluations in an attempt to address the two questions put forth in Chapter 1. Our simulations show that, if units have been properly randomized in the design stage, there is little to separate multiple imputation within and across treatment arms. This of course, depends on the level of imbalance among observed covariates, which in turn depends on the sample size. If n is at least moderate, we can expect that randomization would guarantee good baseline comparability with some minor imbalances existing due to random chance. Imputing using T will add to them, but not enough to warrant highly biased or inefficient estimates. This conclusion however, cannot be generalized to all randomized experiments. Smaller trials (with less than 100 enrolled subjects) are usually prone to larger baseline imbalances. In such cases, we would expect worse estimates if imputing in an arm-dependent way.

If a covariate is imputed using T, it can no longer be deemed independent of the treatment, and this will complicate the interpretation of any causal estimates obtained. The general recommendation would hence be to impute respecting randomization, irrespective of the size of the trial. Relative to imputation using T, this method does not seem to offer massive benefits in terms of bias or precision (at least for large n), but will ensure treatment-covariate independence and hence the validity of causal effects. We also believe that if there is extreme imbalance in the observed covariates, then imputing within categories of T will exacerbate it, but imputing independent of T will retain rather than rectify it. In other words, imputation respecting randomization cannot correct for an improper design.

When the covariates are highly prognostic of the outcome, the results produced by the design and outcome stage methods notably differ. We observe a trade-off between bias and efficiency in this situation. Estimates produced by methods R and NR are highly biased, but have relatively low standard errors. The opposite is true for the outcome stage methods. If the covariate-outcome association is low, all four methods give approximately unbiased estimates, although standard errors are still relatively smaller for R and NR. The design stage methods also mostly overcover, irrespective of the level of association.

To summarize, under the right imputation model, using Y will always be superior to design stage imputation, in terms of producing estimates that are closer to the truth. The definition of going "close to the truth" though, can be subjective. If it means obtaining estimates with minimal bias and the right standard errors, then Y should be a part of the imputation model for X, even if it would imply that the trial design was shaped based on the outcome. If it means getting close to an answer that would have been obtained by strictly following the rules, no matter the bias, then Y (and T) should not be accounted for when imputing the covariates.

Several issues beyond the scope of this thesis exist. First, we have only considered situations where missingness is limited to the covariates. In practice, both the outcome and the covariates have missing values, which certainly complicates inference. In the framework of observed data modeling, it is possible to separate the covariates and the outcome into blocks and place different identifying restrictions within these blocks. For examples of such block-based methods, see Zhou et al. (2010) and Sadinle & Reiter (2018). However, these methods are usually sequential in nature, i.e., any identifying assumption applied to X will in some manner condition on Y. This makes it fairly challenging to incorporate the independence restrictions implied by randomization. Second, we have only considered binary covariates, but we believe that much of what has been concluded will also apply to multinomial and continuous covariates.

In this thesis, we interpret "randomization" to imply that both, the covariates as well as their missingness indicators are independent of the treatment. Even so, situations exist where this independence applies only to the covariates but not to the missingness indicators. If missing values in X occur after treatment assignment, the D become post-treatment variables, whose distribution will not be balanced across treatment arms. An interesting problem would be to identify a full data distribution that allows for the independence of X and T, but not D and T. Lastly, we note that the imputation method we use breaks down each data set by the observed missingness patterns. For this procedure to work well, sufficient number of data points per pattern are required, which amounts to having a large enough sample size. Thus, we did not consider very small-sized experiments, with $n < 100$. We believe that in such trials, chance baseline imbalances will be higher and a clearer difference will be seen between methods respecting and not respecting randomization. Future work on this thesis will look at trials having a smaller sample size, but which is reasonable enough for observed data modeling to work.

Appendix A

Data Generation under ICIN

We begin by slightly modifying Theorem 1 in Sadinle and Reiter (2017) to represent the conditional distribution of $X = x$ given the missingness pattern $D = d$. The η will then be defined recursively as

$$\eta_d(x_{obs}) = \log f_d(x_{obs}) - \log \int_{X_{mis}} \exp \sum_{d^* < d} \eta_{d^*}(x_{obs}) \mathbb{1}(d^* < d)\, dx_{mis},$$

where $f_d(x_{obs}) = f(x_{obs}|d)$. Then,

$$f_d(x) = \exp \sum_{d^* \leqslant d} \eta_{d^*}(x_{obs}).$$

As discussed in Sadinle and Reiter (2017), we can treat each term in η to be a hierarchical loglinear model where the highest order term is a two-way interaction between the observed component of (X_1, X_2) and the indicator of the missing component.

Thus,

$$
\begin{aligned}
\eta_{00}(x_1, x_2) &= \log f(X_1, X_2 | D_1 = 0, D_2 = 0) \\
&= \lambda^{X_1 X_2}_{x_1 x_2} + \lambda^{X_1}_{x_1} + \lambda^{X_2}_{x_2} + \lambda, \\
\eta_{01}(x_1) &= \log f(X_1 | D_1 = 0, D_2 = 1) - \log f(X_1 | D_1 = 0, D_2 = 0) \\
&= \lambda^{X_1 D_2}_{x_1 1} + \lambda^{X_1}_{x_1} + \lambda^{D_2}_{1} + \lambda - \left[\lambda^{X_1}_{x_1} + \lambda\right] \\
&= \lambda^{X_1 D_2}_{x_1 1} + \lambda^{D_2}_{1}, \\
\eta_{10}(x_2) &= \log f(X_2 | D_1 = 1, D_2 = 0) - \log f(X_2 | D_1 = 0, D_2 = 0) \\
&= \lambda^{X_2 D_1}_{x_1 1} + \lambda^{X_2}_{x_2} + \lambda^{D_1}_{1} + \lambda - \left[\lambda^{X_2}_{x_2} + \lambda\right] \\
&= \lambda^{X_2 D_1}_{x_2 1} + \lambda^{D_1}_{1}, \\
\eta_{11} &= \log f(D_1 = 1, D_2 = 1) \\
&\qquad - \log \int_{X_1} \int_{X_2} \exp\left[\eta_{00}(x_1, x_2) + \eta_{01}(x_1) + \eta_{10}(x_2)\right] dx_2 dx_1 \\
&= \lambda^{D_1 D_2}_{11}.
\end{aligned}
$$

The conditional distribution of X given D can be re-expressed in terms of hierarchical log-linear models as

$$
\begin{aligned}
\log f(X_1, X_2 | D_1 = 0, D_2 = 0) &= \eta_{00}(x_1, x_2) \\
&= \lambda + \lambda^{X_1}_{x_1} + \lambda^{X_2}_{x_2} + \lambda^{X_1 X_2}_{x_1 x_2}, \qquad (1) \\
\log f(X_1, X_2 | D_1 = 0, D_2 = 1) &= \eta_{00}(x_1, x_2) + \eta_{01}(x_1) \\
&= \lambda + \lambda^{D_2}_{1} + \lambda^{X_1}_{x_1} + \lambda^{X_2}_{x_2} + \lambda^{X_1 D_2}_{x_1 1} + \lambda^{X_1 X_2}_{x_1 x_2}, \qquad (2)
\end{aligned}
$$

$$
\begin{aligned}
\log f(X_1, X_2 | D_1 = 1, D_2 = 0) &= \eta_{00}(x_1, x_2) + \eta_{10}(x_2) \\
&= \lambda + \lambda_1^{D_1} + \lambda_{x_1}^{X_1} + \lambda_{x_2}^{X_2} + \lambda_{x_2 1}^{X_2 D_1} + \lambda_{x_1 x_2}^{X_1 X_2}, \qquad (3) \\
\log f(X_1, X_2 | D_1 = 1, D_2 = 1) &= \eta_{00}(x_1, x_2) + \eta_{01}(x_1) + \eta_{10}(x_2) + \eta_{11} \\
&= \lambda + \lambda_1^{D_1} + \lambda_1^{D_2} + \lambda_{x_1}^{X_1} + \lambda_{x_2}^{X_2} + \lambda_{x_2 1}^{X_2 D_1} + \lambda_{x_1 1}^{X_1 D_2} \\
&+ \lambda_{x_1 x_2}^{X_1 X_2} + \lambda_{11}^{D_1 D_2}. \qquad (4)
\end{aligned}
$$

Since the covariates and the missingness indicators are all binary, we can treat (1) - (4) as Poisson regression models for counts, with each first order term being the coefficient of an indicator variable taking value 1 if the corresponding variable is 1. The second order terms are coefficients of the products of the first order terms.

Appendix B

Simulation Results under MAR

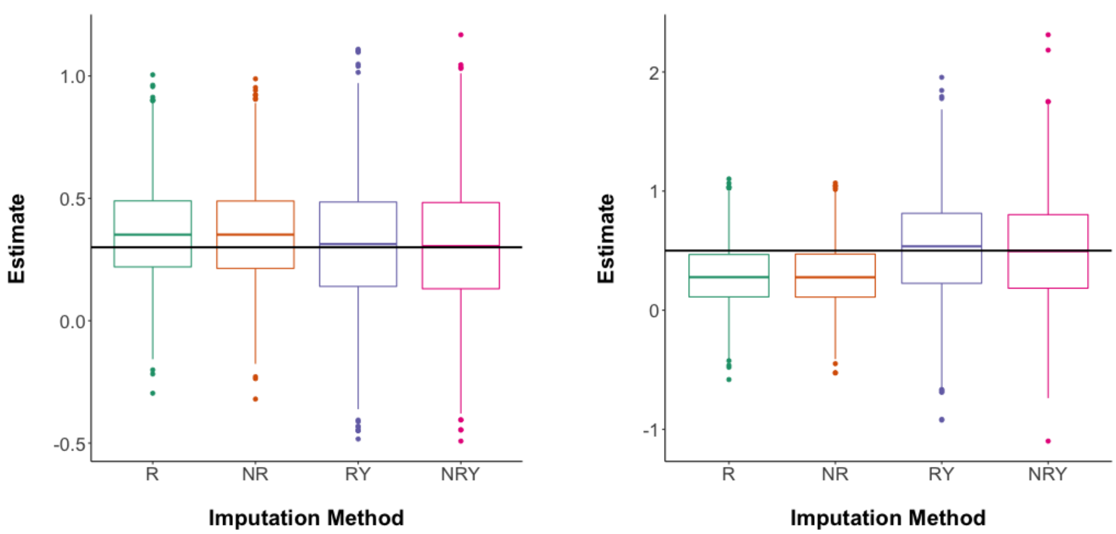

FIGURE B.1: The left panel represents the distribution of the β_t estimates over 1000 replications, where the true value of $\beta_t = 0.3$. The right panel represents the distribution of the β_{tx_2} estimates over 1000 replications, where the true value of $\beta_{tx_2} = 0.5$. Estimates for both coefficients have been obtained under MAR for the S_{Ind}, high association setting.

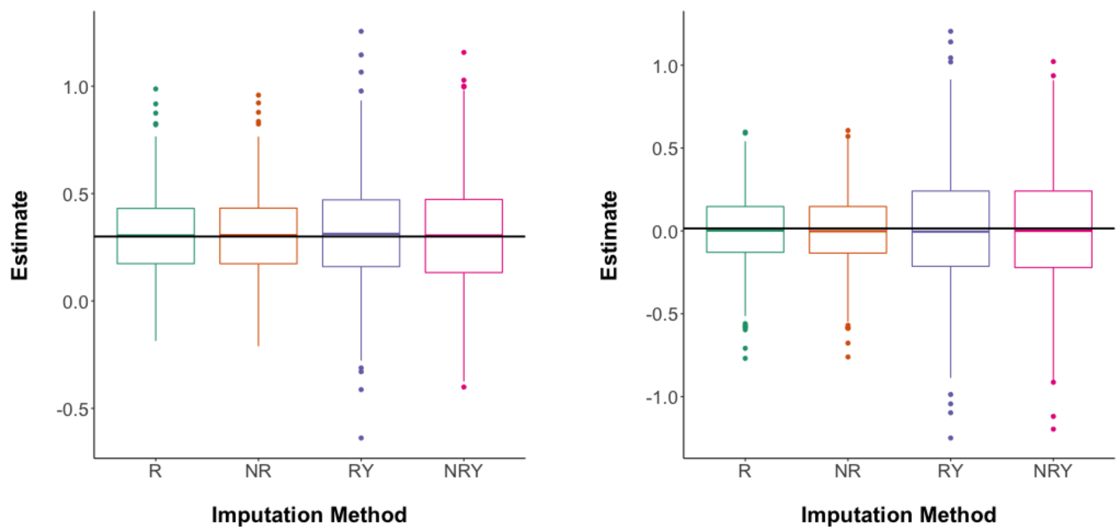

FIGURE B.2: The left panel represents the distribution of the β_t estimates over 1000 replications, where the true value of $\beta_t = 0.3$. The right panel represents the distribution of the β_{tx_2} estimates over 1000 replications, where the true value of $\beta_{tx_2} = 0.015$. Estimates for both coefficients have been obtained under MAR for the S_{Ind}, low association setting.

Table B.1: Standard errors and coverage rates for β_t and β_{tx_2} under the S_{Ind}, high association setting.

	β_t			β_{tx_2}		
Method	MC-SD	MI-SE	Coverage	MC-SD	MI-SE	Coverage
R	0.2034	0.2449	0.976	0.245	0.3770	0.979
NR	0.2037	0.2500	0.981	0.2429	0.3792	0.98
RY	0.2467	0.2515	0.949	0.424	0.4445	0.942
NRY	0.2587	0.2530	0.943	0.4591	0.4411	0.938

Table B.2: Standard errors and coverage rates for β_t and β_{tx_2} under the S_{Ind}, low association setting.

	β_t			β_{tx_2}		
Method	MC-SD	MI-SE	Coverage	MC-SD	MI-SE	Coverage
R	0.1812	0.2243	0.986	0.2090	0.3077	0.998
NR	0.1815	0.2244	0.985	0.2091	0.3078	0.997
RY	0.2364	0.2380	0.953	0.3363	0.3341	0.953
NRY	0.2463	0.2385	0.944	0.3340	0.3361	0.949

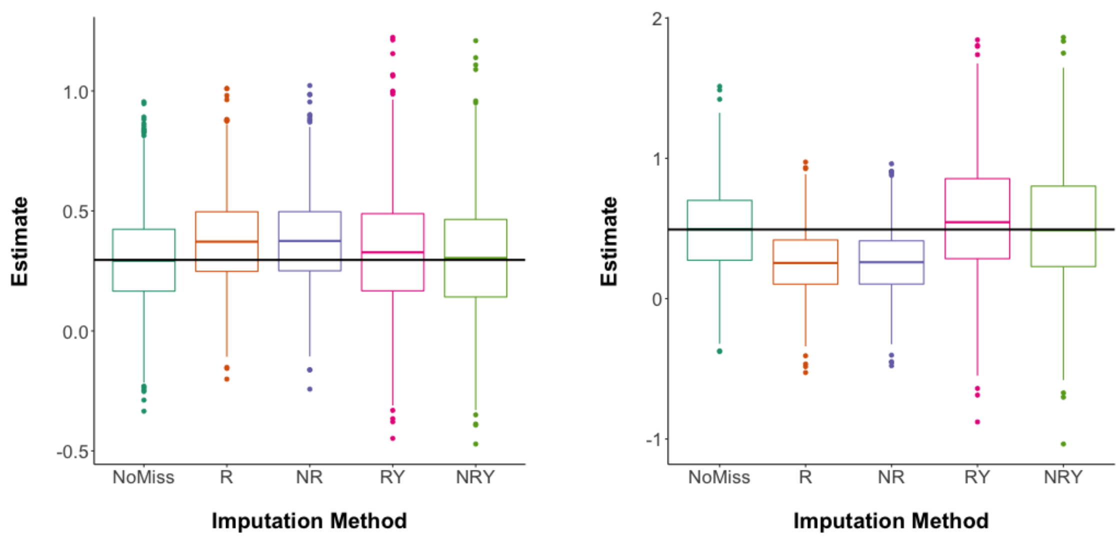

FIGURE B.3: The left panel represents the distribution of the β_t estimates over 1000 replications, where the true value of $\beta_t = 0.3$. The right panel represents the distribution of the β_{tx_2} estimates over 1000 replications, where the true value of $\beta_{tx_2} = 0.5$. Estimates for both coefficients have been obtained under MAR for the S_T, high association setting.

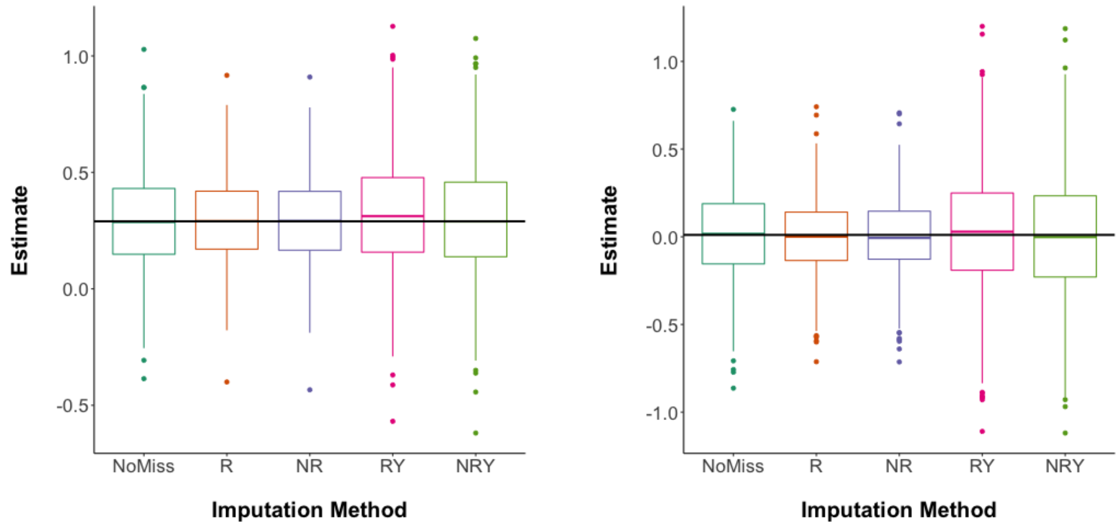

FIGURE B.4: The left panel represents the distribution of the β_t estimates over 1000 replications, where the true value of $\beta_t = 0.3$. The right panel represents the distribution of the β_{tx_2} estimates over 1000 replications, where the true value of $\beta_{tx_2} = 0.015$. Estimates for both coefficients have been obtained under MAR for the S_T, low association setting.

Table B.3: Standard errors and coverage rates for β_t and β_{tx_2} under the S_T, high association setting.

	β_t			β_{tx_2}		
Method	MC-SD	MI-SE	Coverage	MC-SD	MI-SE	Coverage
R	0.1878	0.2392	0.978	0.2299	0.3522	0.978
NR	0.1884	0.2395	0.982	0.2299	0.3527	0.98
RY	0.2457	0.2468	0.951	0.4116	0.4045	0.946
NRY	0.2443	0.2448	0.95	0.4207	0.4116	0.955
NoMiss	0.2036	0.2095	0.95	0.3022	0.3036	0.95

Table B.4: Standard errors and coverage rates for β_t and β_{tx_2} under the S_T, low association setting.

	β_t			β_{tx_2}		
Method	MC-SD	MI-SE	Coverage	MC-SD	MI-SE	Coverage
R	0.1783	0.2233	0.989	0.2093	0.3063	0.996
NR	0.1795	0.2040	0.986	0.2115	0.3073	0.994
RY	0.2412	0.2362	0.942	0.3390	0.3311	0.948
NRY	0.2412	0.2382	0.949	0.343	0.3356	0.942
NoMiss	0.2029	0.2019	0.949	0.2599	0.2460	0.95

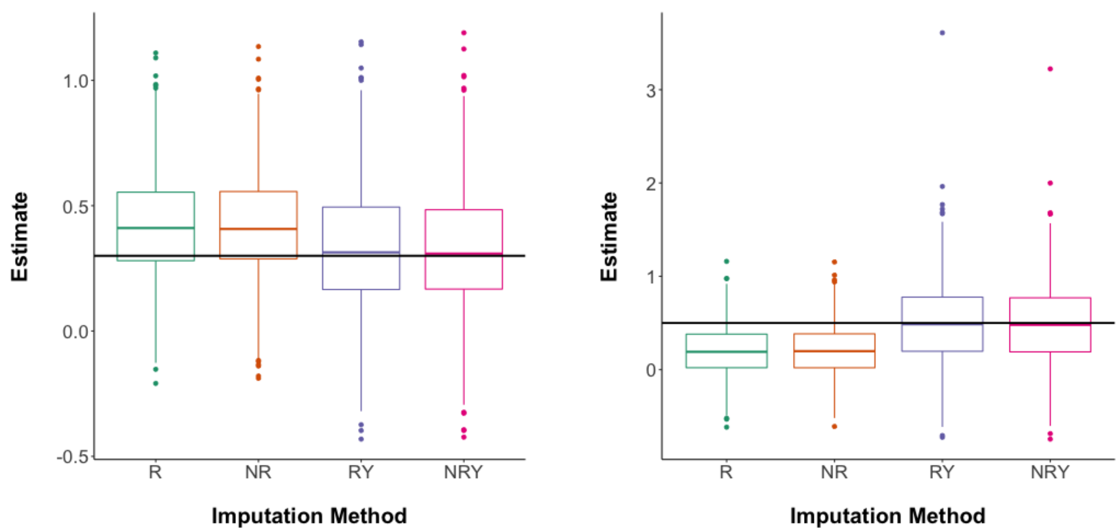

FIGURE B.5: The left panel represents the distribution of the β_t estimates over 1000 replications, where the true value of $\beta_t = 0.3$. The right panel represents the distribution of the β_{tx_2} estimates over 1000 replications, where the true value of $\beta_{tx_2} = 0.5$. Estimates for both coefficients have been obtained under MAR for the S_Y, high association setting.

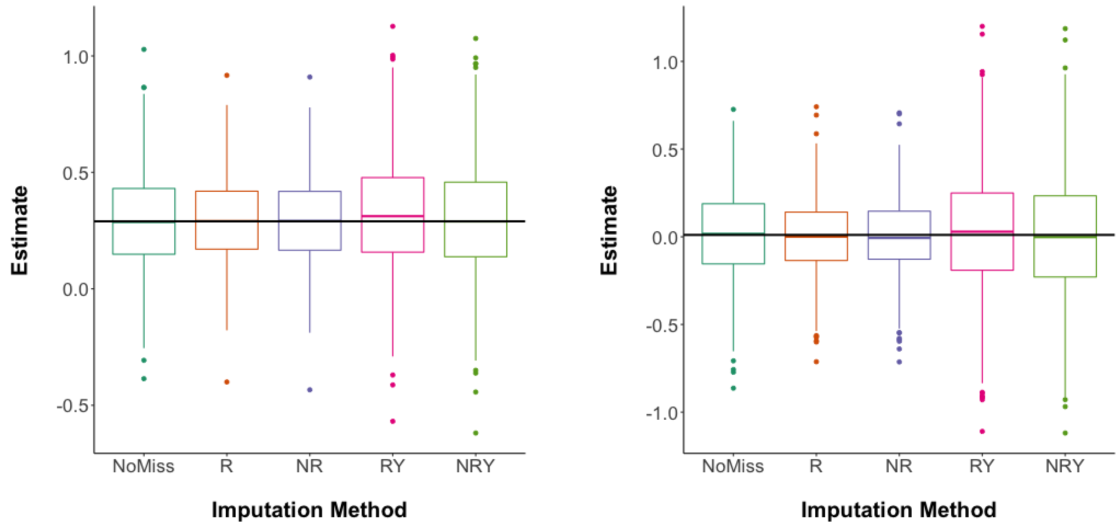

FIGURE B.6: The left panel represents the distribution of the β_t estimates over 1000 replications, where the true value of $\beta_t = 0.3$. The right panel represents the distribution of the β_{tx_2} estimates over 1000 replications, where the true value of $\beta_{tx_2} = 0.015$. Estimates for both coefficients have been obtained under MAR for the S_Y, low association setting.

Table B.5: Standard errors and coverage rates for β_t and β_{tx_2} under the S_Y, high association setting.

	β_t			β_{tx_2}		
Method	MC-SD	MI-SE	Coverage	MC-SD	MI-SE	Coverage
R	0.2016	0.2436	0.965	0.2649	0.3729	0.951
NR	0.2023	0.2422	0.964	0.2663	0.3745	0.953
RY	0.2426	0.2429	0.946	0.4415	0.4286	0.943
NRY	0.2435	0.2438	0.949	0.4412	0.4281	0.958

Table B.6: Standard errors and coverage rates for β_t and β_{tx_2} under the S_Y, low association setting.

	β_t			β_{tx_2}		
Method	MC-SD	MI-SE	Coverage	MC-SD	MI-SE	Coverage
R	0.1816	0.2210	0.980	0.2012	0.3012	0.993
NR	0.1757	0.2185	0.982	0.2088	0.3015	0.952
RY	0.2134	0.2286	0.942	0.3184	0.3321	0.952
NRY	0.2391	0.2345	0.954	0.3233	0.3091	0.946

Bibliography

Agresti, A. (2012), *Categorical Data Analysis*, Hoboken, NJ: John Wiley & Sons.

Allison, P. D. (2002), *Quantitative Applications in the Social Sciences: Missing Data*, Thousand Oaks, CA: SAGE Publications, Inc.

Barnard, J. and Meng, X. L. (1999), "Application of multiple imputation in medical studies: from AIDS to NHANES," *Statistical Methods in Medical Research*, 8, 17–36.

Birmingham, J., Rotnitzky, A., and Fitzmaurice, G. M. (2003), "Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 275–297.

Chen, H., Geng, Z., and Zhou, X. H. (2009), "Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data (with discussion)," *Biometrics*, 65(3), 675–682.

Daniels, M. J. and Hogan, J. W. (2008), *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*, Boca Raton, FL: Chapman and Hall/CRC.

Ding, P. and Geng, Z. (2014), "Identifiability of subgroup causal effects in randomized experiments with nonignorable missing covariates," *Statistics in Medicine*, 33, 1121–1133.

Frangakis, C. E. and Rubin, D. B. (1999), "Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes," *Biometrika*, 86(2), 365–379.

Gill, R. D., van der Laan, M. J., and Robins, J. M. (1997), "Coarsening at Random: Characterizations, Conjectures, Counter-Examples," in *Proceedings of the First Seattle Symposium in Biostatistics*, pp. 255–294.

Greenland, S. and Finkle, W. D. (1995), "A critical look at methods for handling missing covariates in epidemiologic regression analyses," *American Journal of Epidemiology*, 142, 1255–1264.

Harel, O. and Zhou, X. H. (2007), "Multiple imputation: review of theory, implementation and software," *Statistics in Medicine*, 26, 3057–3077.

Horowitz, J. L. and Manski, C. F. (2000), "Nonparametric analysis of randomized experiments with missing covariate and outcome data (with discussion)," *Journal of the American Statistical Association*, 95(449), 77–84.

Imai, K. (2009), "Statistical analysis of randomized experiments with non-ignorable missing binary outcomes: An application to a voting experiment," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(1), 83–104.

Imbens, G. W. and Pizer, W. A. (2000), "The analysis of randomized experiments with missing data," *Resources for the Future, Discussion Paper*, pp. 00–19.

Kahan, B. C., Jairath, V., Dore, C. J., and Morris, T. P. (2014), "The risks and rewards of covariate adjustment in randomized trials: An assessment of 12 outcomes from 8 studies," *Trials*, 15(139), 1–7.

Linero, A. R. (2017), "Bayesian nonparametric analysis of longitudinal studies in the presence of informative missingness," *Biometrika*, 104, 327–341.

Linero, A. R. and Daniels, M. J. (2018), "Bayesian approaches for missing not at random outcome data: The role of identifying restrictions," *Statistical Science*, 33, 198–213.

Little, R. J. A. (1992), "Regression with missing X's: A review," *Journal of the American Statistical Association*, 87(420), 1227–1237.

Little, R. J. A. (1993), "Pattern-mixture models for multivariate incomplete data," *Journal of the American Statistical Association*, 88, 125–134.

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

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

Polson, N. G., Scott, J. G., and Windle, J. (2013), "Bayesian inference for logistic models using polya-gamma latent variables," *Journal of the American Statistical Association*, 108, 1339–1349.

Reiter, J. P. and Raghunathan, T. E. (2007), "The multiple adaptations of multiple imputation," *Journal of the American Statistical Association*, 102, 1462–1471.

Robins, J. M. (1997), "Non-response models for the analysis of non-monotone non-ignorable missing data," *Statistics in Medicine*, 16(1), 21–37.

Robins, J. M. and Wang, N. (2000), "Inference for imputation estimators," *Biometrika*, 87, 113–124.

Robinson, L. D. and Jewell, N. P. (1991), "Some surprising results about covariate adjustment in logistic regression models," *International Statistical Review*, 58, 227–240.

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

Rubin, D. B. (1978), "Bayesian inference for causal effects: The role of randomization," *Annals of Statistics*, 6(1), 34–58.

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

Rubin, D. B. (1996), "Multiple imputation after 18+ years," *Journal of the American Statistical Association*, 91(434), 473–489.

Rubin, D. B. (2002), "Using propensity scores to help design observational studies: Application to the Tobacco Litigation," *Health Services and Outcomes Research Methodology*, 2, 169–188.

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, 20–30.

Rubin, D. B. (2008), "For objective causal inference, design trumps analysis," *Annals of Applied Statistics*, 2 (3), 808–840.

Rubin, D. B. and Schenker, N. (1991), "Multiple imputation in health-care databases: An overview and some applications," *Statistics in Medicine*, 10, 585–598.

Sadinle, M. and Reiter, J. P. (2017), "Itemwise conditionally independent nonresponse modelling for incomplete multivariate data," *Biometrika*, 104(1), 207–220.

Sadinle, M. and Reiter, J. P. (2018), "Sequential identification of nonignorable missing data mechanisms," *Statistica Sinica*, 28, 1741–1759.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman Hill.

Scharfstein, D., Onicescu, G., Goodman, S., and Whitaker, R. (2011), "Analysis of subgroup effects in randomized trials when subgroup membership is missing: Application to the second multicenter automatic defibrillator intervention trial," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60 (4), 607–617.

Schemper, M. and Smith, T. L. (1990), "Efficient evaluation of treatment effects in the presence of missing covariate values," *Statistics in Medicine*, 9, 777–784.

Schulz, K. F. and Grimes, D. A. (2002), "Sample size slippages in randomized trials: Exclusions and the lost and wayward," *Lancet*, 359, 781–785.

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

Sullivan, T. R., White, I. R., Salter, A. B., Ryan, P., and Lee, K. J. (2018), "Should multiple imputation be the method of choice for handling missing data in randomized trials?" *Statistical Methods in Medical Research*, 27(9), 2610–2626.

Tchetgen Tchetgen, E. J., Wang, L., and Sun, B. (2016), "Discrete choice models for nonmonotone nonignorable missing data: Identification and inference," Preprint. Available at arXiv:1607.02631.

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., and Curran, D. (2002), "Strategies to fit pattern-mixture models," *Biostatistics*, 32, 245–265.

Vach, W. (1994), *Logistic Regression with Missing Values in the Covariates*, New York, NY: Springer.

Vach, W. and Blettner, M. (1991), "Biased estimates of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variable," *American Journal of Epidemiology*, 134, 895–907.

Vansteelandt, S. R., Goetghebeur, E., Kenward, M., and Molenberghs, G. (2006), "Ignorance and uncertainty regions as inferential tools in a sensitivity analysis," *Statistica Sinica*, 16, 953–979.

Vansteelandt, S. R., Rotnitzky, A., and Robins, J. M. (2007), "Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse," *Biometrika*, 94, 841–860.

Wang, N. and Robins, J. M. (1998), "Large-sample theory for parametric multiple imputation procedures," *Biometrika*, 85(4), 935–948.

White, I. R. and Thompson, S. G. (2005), "Adjusting for partially missing baseline measurements in randomized trials," *Statistics in Medicine*, 24, 993–1007.

Zhou, Y., Little, R. J. A., and Kalbfleisch, J. D. (2010), "Block-conditional missing at random models for missing data," *Statistical Science*, 25, 517–532.