

**Statistical Approaches for Missing Covariates and a Novel Joint Model for
Ordered Bivariate Survival Times**

by

Jiaheng Xie

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2025

Doctoral Committee:

Professor Alexander Tsodikov, Co-Chair
Professor Min Zhang, Co-Chair
Associate Professor Kevin He
Professor David Mendez

Jiaheng Xie

jiahengx@umich.edu

ORCID iD: 0000-0001-6466-5616

© Jiaheng Xie 2025

ACKNOWLEDGEMENTS

It is hard to believe that my graduate school journey is coming to an end. These six years in Ann Arbor have been truly remarkable, and I am deeply grateful to all the people who made this experience so special.

First and foremost, I would like to express my heartfelt gratitude to my advisors, Dr. Alexander Tsodikov and Dr. Min Zhang. Our weekly meetings were a constant source of inspiration and intellectual growth. Dr. Tsodikov, your infectious passion for research has been a driving force throughout my journey. I am particularly thankful for your unwavering support; whenever I faced challenges in my research, you were always available for a quick meeting to guide me forward. Dr. Zhang's thoughtful and methodical approach to problem-solving, which emphasizes a deep understanding of context before addressing complex questions, has profoundly influenced my research methodology. Dr. Zhang's exceptional dedication to my work has been truly inspiring. Your meticulous attention to detail, commitment to clarity and precision, and exceptional communication skills have deeply impacted my scientific thinking. Their combined expertise, humility, curiosity, and passion for discovery have not only shaped my dissertation but have also instilled in me the values of a conscientious and responsible researcher. I will forever hold their mentorship in the highest regard, and they will always have my deepest admiration and respect.

I also wish to extend my sincere gratitude to the members of my committee, Dr. Kevin He and Dr. David Mendez. Dr. He's enthusiasm for statistics ignited my own passion for the field, and Dr. Mendez's unwavering support and detailed feedback on my dissertation proposal were invaluable.

I am deeply appreciative of my dedicated colleagues in the lab. Their unwavering support, insightful discussions, and collaborative spirit have been instrumental to the success of my project. I am especially thankful for their willingness to share their expertise and provide encouragement throughout this endeavor.

I feel incredibly fortunate to have been surrounded by a supportive network of classmates and friends during my time in Ann Arbor. Over the past seven years, their invaluable assistance, thoughtful suggestions, and camaraderie have greatly enriched my graduate school experience.

Finally, I would like to express my profound gratitude to my parents. Their unwavering belief in me and unconditional love have been a constant source of strength and motivation. Despite the physical distance, their support has been instrumental in all my accomplishments and has meant the world to me.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	ii
LIST OF FIGURES	vi
LIST OF TABLES	vii
LIST OF APPENDICES	viii
ABSTRACT	ix
CHAPTER	
1 Introduction	1
2 Robust Covariate Adjustment for Randomized Clinical Trials when Co- variates are Subject to Missingness	3
2.1 Introduction	3
2.2 Notation	5
2.3 Background Information	6
2.3.1 Robust covariate adjustment framework	6
2.3.2 Methods for handling missing covariates	7
2.4 Method	8
2.4.1 Complete case analysis	8
2.4.2 Imputation method	9
2.4.3 General augmented method	10
2.5 Simulation Studies	11
2.5.1 Univariate missingness	12
2.5.2 Multivariate missingness	14
2.5.3 Multivariate missingness with more covariates	16
2.5.4 Simulation based on ACTG 175 trial	16
2.6 Application	18
2.7 Discussion	19
3 Cautious Notes of Multiple Imputation in Survival Analysis: a Simulation Study	21
3.1 Introduction	21
3.2 Notations and missing data method	23

3.2.1	Notations	23
3.2.2	Missing data methods in survival analysis	23
3.3	Missing data mechanisms	26
3.4	Simulation	28
3.4.1	Simulation 1: two covariates, univariate missingness	29
3.4.2	Simulation 2: multivariate missingness	32
3.4.3	Summary	32
3.5	Discussion	35
4	Joint Models for Ordered Bivariate Survival times on Semi-Competing Risk Data	36
4.1	Introduction	36
4.2	Notation and Data Structure	37
4.3	Modified Copula and Model Specification	38
4.3.1	Copula With Order Restriction	38
4.3.2	Model Specification	39
4.4	Likelihood and Estimation	40
4.4.1	Likelihood Construction	40
4.4.2	Estimation	41
4.5	Simulation Studies	45
4.5.1	Finite-sample Properties of Parameter Estimates	45
4.5.2	Hypothesis Testing	46
4.6	Data Application	46
4.7	Discussion	47
5	Conclusion	49
	APPENDICES	51
	BIBLIOGRAPHY	69

LIST OF FIGURES

FIGURE

3.1	PB and MC MSE of X_1 in simulation 1. Dashed lines represent the $\pm 10\%$ PB.	30
3.2	SD and Cov. prob. of X_1 in simulation 1. Dashed lines represent the 95% and 80% coverage probability.	31
3.3	PB and MC MSE of X_1 in simulation 2. Dashed lines represent the $\pm 10\%$ PB.	33
3.4	SD and Cov. prob. of X_1 in simulation 2. Dashed lines represent the 95% and 80% coverage probability.	34
B.1	PB and MC MSE of Z_1 in simulation 1. Dashed lines represent the $\pm 10\%$ PB. .	58
B.2	SD and Cov. prob. of Z_1 in simulation 1. Dashed lines represent the 95% and 80% coverage probability.	59
B.3	PB and MC MSE of Z_1 in simulation 2. Dashed lines represent the $\pm 10\%$ PB. .	60
B.4	SD and Cov. prob. of Z_1 in simulation 2. Dashed lines represent the 95% and 80% coverage probability.	61

LIST OF TABLES

TABLE

2.1	Simulation results for univariate missingness based on 1000 Monte Carlo datasets, with a sample size of 800.	13
2.2	Simulation results for multivariate missingness based on 1000 Monte Carlo datasets when data are MNAR, with a sample size of 800.	15
2.3	Simulation results for 30 missing covariates based on 1000 Monte Carlo datasets, with a sample size of 800.	17
2.4	Simulation results based on the ACTG data with 1000 Monte Carlo datasets and a sample size of 2000.	18
2.5	Data structure for ACTG data	19
2.6	Estimates of β for the ACTG data, relative efficiency is calculated by standard error only	20
3.1	Imputation methods	25
3.2	Missing mechanisms.	28
4.1	Simulation results for the proposed model	46
4.2	Estimations from the proposed model and Cox model for cancer death	47
B.1	Missing mechanisms, percentages, and corresponding missing models.	55
B.2	Missing mechanisms, percentages, and corresponding missing models.	56
B.3	Missing mechanisms, percentages, and corresponding missing models.	57

LIST OF APPENDICES

A Appendix Chapter 2	51
Equivalence between the estimand and expectation difference of the outcome . . .	51
Consistency of estimand using different imputation methods	51
Imputation as a two-step augmented method	52
Comparison of imputation method and modified imputation method	52
The asymptotic equivalence of ANCOVA1 and the augmented method	53
B Appendix Chapter 3	55
Simulation 1: Missing Mechanisms	55
Simulation 2: Missing Mechanisms	56
Simulation 2: Missing Mechanisms Continued	57
Figures	58
C Appendix Chapter 4	62
Joint survival function in toy example and its corresponding copula	62
Model restrictions to satisfy the order restriction of cancer diagnosis and cancer death	62
No lambda information contained in data	64
EM algorithm	64

ABSTRACT

This dissertation develops and evaluates methods for some common statistical challenges arising from clinical and epidemiological studies.

Missing covariates is a very common problem in clinical and epidemiologic studies and may lead to biased estimates and inference if we do not account for missingness. Chapters II and III study the problem of missing covariates. Specifically, Chapter II focuses on estimating the causal treatment effect in randomized clinical trials when covariate adjustment is used for improving efficacy, but covariates are possibly missing. We adopt the robust augmentation framework for covariate adjustment and propose three different implementation strategies when multiple baseline covariates are likely missing. Without assuming the data are MAR, the proposed methods lead to consistent and asymptotically normal estimators of the treatment effects, as long as the treatment assignment is independent of the missingness. Chapter III focuses on handling missing covariates in the Cox regression model, which is challenging as the outcome is subject to censoring. Existing methods generally require MAR, but this assumption may not be plausible in survival analysis. Instead of MAR, two alternative missing assumptions are proposed. However, the performance of commonly used imputation methods in this context is relatively less studied. We conducted extensive simulation studies to evaluate the performance of some popular methods, for example, the multiple imputation by chained equations (MICE), complete-case analysis (CC), and single imputation (SI), under various missingness mechanisms and levels of missingness.

Chapter IV develops methods for analyzing ordered bivariate survival times, motivated by the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, to study the effect of screening on cancer mortality. We aim to jointly model the time to cancer diagnosis and the time to death to improve efficiency. Our joint model accounts for the ordering of events, where cancer diagnosis must occur before death. A copula is a function that links two random variables by specifying their dependence structure. Copulas have been extensively applied in bivariate survival models. However, traditional copulas do not account for the order restrictions between random variables. To address this limitation, we propose a novel joint model for ordered bivariate survival times based on copulas.

CHAPTER 1

Introduction

In Chapter II, we develop statistical methods for robust covariate adjustment to improve efficiency of inferences in randomized clinical trials when covariates are subject to missingness. In randomized clinical trials, often the primary goal is to estimate the treatment effect. Robust covariate adjustment is a preferred statistical method since it improves efficiency and is robust to model misspecification. However, it is still underutilized in practice. One practical challenge is the missing covariates. While extensive literature on missing covariates exists, most methods have been developed in the regression setting, primarily focusing on the estimation of regression coefficients. Furthermore, these methods often require additional assumptions about the missingness mechanism, such as the missing at random (MAR) assumption. In randomized clinical trials, however, the primary goal is to estimate the causal treatment effect. Covariates can be leveraged to improve efficiency, but their association with outcomes is treated as auxiliary information. Moreover, methods that rely on strong assumptions about missingness mechanisms are less appealing. Therefore, existing methods for handling missing covariates cannot be directly applied. In this chapter, we adopt the robust augmentation framework for covariate adjustment and propose three different implementation strategies for scenarios where multiple baseline covariates may be missing. We evaluate the performance of the proposed methods through extensive simulations and demonstrate their practical utility by applying them to data from the AIDS Clinical Trials Group (ACTG) protocol 175.

In Chapter III, we study the problem of covariates missingness in regression analysis for survival outcomes using the Cox model. Several approaches have been developed for handling missing covariates in survival settings. These methods typically require that data are missing at random (MAR), i.e., the probability of missingness depends only on the observed data. Also, the methods' performances are usually evaluated under the assumption of MAR. However, in survival analysis, because the outcome is subject to censoring, the MAR assumption may not be plausible and is hard to justify when the missingness depends on the

outcome. Alternative and arguably more plausible missingness assumptions have been proposed, such as censoring ignorable missing at random (CIMAR) and failure-ignorable missing at random (FIMAR). There is a lack of simulation studies evaluating the performance of popular methods for handling missing covariates in survival analysis under various missingness mechanisms. To address this limitation, We conducted extensive simulation studies to study the performance of some popular methods, for example, the multiple imputation by chained equations (MICE), complete-case analysis (CC), and single imputation (SI), under various missingness mechanisms and levels of missingness.

In Chapter IV, we develop a novel joint model for ordered bivariate survival times of cancer diagnosis and cancer death to evaluate the screening effect on mortality using copulas. A copula is a bivariate distribution function that links two random variables via their marginal distributions by specifying the dependence structure. While extensive literature exists on copulas, traditional copulas do not guarantee the order restriction of random variables. In our case, cancer diagnosis must precede cancer death. There is a lack of research on copulas that explicitly consider order restrictions. In this chapter, We extend traditional copulas to accommodate scenarios with order restrictions. The extended copulas are then used to construct the joint likelihood for ordered cancer diagnosis and cancer death. This framework is applied to assess the screening effect on mortality, with the goal of improving efficiency. We evaluate the performance of the proposed methods through simulation studies and real-world data application.

CHAPTER 2

Robust Covariate Adjustment for Randomized Clinical Trials when Covariates are Subject to Missingness

2.1 Introduction

Generally, in randomized clinical trials, the primary goal is to compare the average treatment effect (ATE) among treatment groups. Along with the primary outcome and treatment assignment, extensive baseline variables are collected before assigning treatment. Statisticians have reached a consensus (Assmann et al., 2000; Pocock et al., 2002; Food et al., 2020; Benkeser et al., 2021) that one may obtain a more efficient estimator by appropriately adjusting for prespecified, prognostic baseline variables (i.e. variables that are correlated with the outcomes). This adjustment is known as covariate adjustment. To leverage information from baseline variables and be robust to model misspecification at the same time, methods for robust covariate adjustment have been developed. There is extensive literature on this topic.(Yang and Tsiatis, 2001; Tsiatis et al., 2008; Zhang et al., 2008; Moore and van der Laan, 2009; Jiang et al., 2019) In particular, Tsiatis et al.Tsiatis et al. (2008) and Zhang et al.Zhang et al. (2008) exploited semiparametric theory and studied influence functions of all consistent and asymptotically normal estimators for quantifying treatment effect under unrestrictive assumptions. Valid inference and the improvement of efficiency hold even if the working model for the relationship between the outcome and covariates is misspecified. In clinical trials, however, robust covariate adjustment is still underused in practice. One challenge limiting its use is the missingness in covariates. A complete-case analysis on patients with complete covariates may lose efficiency relative to the unadjusted analysis due to the smaller sample size. More seriously, it may introduce bias due to the imbalanced selection of patients. Therefore, it is important to study methods for robust covariate adjustment in randomized clinical trials when covariates are likely missing to promote the use of covariate

adjustment in practice.

Missing covariates is a common issue in analysis, not unique to the setting of randomized clinical trials. A large body of literature on missing covariates in the regression setting exists. We refer readers to Little (1992) and Ibrahim et al. (2005) for literature reviews and Little and Rubin, Little and Rubin (2019) Schafer (1997) for further detailed studies. It is well known that complete case analysis is prone to bias. Other common methods for handling missing covariates have been developed, which include single imputation, (Little, 1992) maximum likelihood, (Rubin, 1976) multiple imputation, (Rubin, 2018) inverse probability weighting (IPW) (Robins et al., 1994) and fully bayesian methods. (Ibrahim et al., 2005) However, most of the methods mentioned above focus on estimating regression models and studying the relationship between outcome and regressors (covariates). In addition, those methods often require additional assumptions (e.g., missing at random in multiple imputation). For these reasons, the methods discussed above may not be optimal for our purpose. This is because in our setting, we aim to consistently estimate the treatment effect and the relationship between covariates and the outcomes is a nuisance exploited to improve efficiency, which is in contrast to regression. Moreover, since the unadjusted method, this benchmark is already robust in randomized clinical trials, robustness is crucial and methods that require additional assumptions on missingness are not appealing.

To estimate treatment effect with missing covariates in randomized clinical trials, several methods such as unadjusted analysis, single imputation, (Kayembe et al., 2020) missing indicator method (Groenwold et al., 2012) and multiple imputation (Rubin, 1976; Sullivan et al., 2018) have been evaluated through simulation studies. However, these studies lack a theoretical framework of robust covariate adjustment on missing covariates. To the best of our knowledge, little work (Zhao and Ding, 2022; Chang et al., 2022) has been done on missing covariates in the context of robust covariate adjustment for estimating treatment effect. Moreover, only one covariate subjecting to missingness is uncommon in practice. Most often, multiple baseline covariates may be missing, which could complicate the working model. However, most of the literature discussed above imposed missingness on only one covariate and didn't consider multiple incomplete covariates. Thus there is a dearth of guidance for handling multivariate missing covariates using robust covariate adjustment framework.

In this article, we aim to study the robust covariate adjustment on estimating the treatment effect in clinical trials when multivariate baseline covariates are missing. For the purpose of simplicity, we assume that the outcome is continuous and fully observed throughout this article. However, general results could be extended for binary, ordinal, or other types of outcomes. In Section 2.2, we introduce basic notations and the estimand of interest. Section 2.3.1 reviews the robust covariate adjustment framework and its practical imple-

mentation, based on Zhang’s work. Zhang et al. (2008) Section 2.3.2 reviews methods for handling missing covariates in clinical trials. In Section 2.4, we propose several methods for handling missing covariates integrated with the robust covariate adjustment. In Section 2.5, we evaluate the performance of our proposed strategies in different missing scenarios. Real data application is shown in Section 2.6 and a brief discussion is presented in Section 2.7.

2.2 Notation

We assume that $(Y_i, \mathbf{X}_i, A_i), i = 1, \dots, n$, are independent and identically distributed from a k -arm randomized clinical trial, $k \geq 2$. Let Y_i denote the continuous outcome, $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$ denotes the auxiliary baseline covariates which might be missing, with $p \times 1$ dimension. We focus on $k = 2$ in this paper, use $A_i = 0$ to indicate assigning to control, and $A_i = 1$ to indicate assigning to treatment. We define π as the probability of getting treatment. Let $Y_i^{(a)}, a = 0, 1$ denote the potential outcome if assigned to control or treatment respectively. Let $\mathbf{R}_i = (R_{i1}, \dots, R_{ip})^T$ be a $p \times 1$ observation indicator vector for $\mathbf{X}_i$, such that $R_{ij} = 0$ if X_{ij} is missing and $R_{ij} = 1$ if X_{ij} is observed, $j = 1, \dots, p$.

Suppose the estimand we are interested in is the Average Treatment Effect (ATE),

$$\theta = E(Y_i^1) - E(Y_i^0). \quad (2.1)$$

In randomized clinical trials, it has been proved that ATE is equivalent to the difference in expectations of Y between treatment and control under the stable unit treatment value assumption (see Appendix A.1). A natural and commonly used estimator for ATE is then the unadjusted estimator, constructed by the difference of sample means in the treatment group and control group.

$$\hat{\theta}_{un} = \frac{\sum_{i=1}^n A_i Y_i}{\sum_{i=1}^n A_i} - \frac{\sum_{i=1}^n (1 - A_i) Y_i}{\sum_{i=1}^n (1 - A_i)} = \bar{Y}_1 - \bar{Y}_0. \quad (2.2)$$

This estimator is consistent, easy to implement, and does not rely on any model assumptions. However, the efficiency could be improved since the unadjusted estimator ignores the baseline information.

With the correlation between baseline covariates and the outcome, a more efficient estimator can be obtained through robust covariate adjustment. This estimator is not only robust to model misspecification and consistent with our estimand θ but also more efficient than the unadjusted estimator by leveraging the baseline information. However, robust covariate

adjustment is still underutilized in practice. One of the critical reasons is the multivariate missing baseline covariates, which is a common issue in reality and needs to be addressed.

2.3 Background Information

2.3.1 Robust covariate adjustment framework

Here we give a general review of the covariate adjustment framework and a practical strategy for analysis based on Zhang. Zhang et al. (2008) Let $(Y_i, \mathbf{X}_i, A_i), i = 1, \dots, n$, be i.i.d. In fact, the outcome Y could be continuous, binary, ordinal, or other types while we focus on continuous in this paper. The class of consistent and asymptotically normal estimators for θ are solutions to

$$m^*(Y, \mathbf{X}, A, \theta) = m(Y, A, \theta) - (A - \pi)h(\mathbf{X}), \quad (2.3)$$

where $m(Y, A, \theta)$ is any unbiased estimating function, $h(\mathbf{X})$ is an arbitrary function of $\mathbf{X}$. The optimal $h(\mathbf{X})$, which leads to the estimator with the smallest variance within the class, is defined as

$$h(\mathbf{X}) = E(m(Y, A, \theta)|A = 1, \mathbf{X}) - E(m(Y, A, \theta)|A = 0, \mathbf{X}). \quad (2.4)$$

The optimal estimator in (2.3) depends on conditional expectations $E(m(Y, A, \theta)|A = a, \mathbf{X}), a = 0, 1$, which need to be estimated because we do not know the true form. Result (2.4) suggests that the closer the working model to the true form, the better the efficiency is achieved. When $h(\mathbf{X}) = 0$, solving (2.3) leads to the unadjusted estimator. It may seem that the efficiency gain of the optimal estimator over the unadjusted one relies on the correct modeling of (2.4). However, suppose we build linear regression working models as $E(m(Y, A, \theta)|A = a, \mathbf{X}) = \kappa_{\mathbf{a}}^T g(\mathbf{X}), a = 0, 1$ and fit the above models via ordinary least square (OLS), where $g(\mathbf{X})$ is a vector of basis function of $\mathbf{X}$ and $\kappa_{\mathbf{a}}$ is the vector of corresponding coefficients. This augmented estimator, obtained by the likely misspecified models, is guaranteed to be asymptotically equal or more efficient than the unadjusted estimator by the argument similar in Leon et al. Leon et al. (2003) Therefore, the efficiency gain does not rely on correct modeling and is robust to working model misspecification. Another important fact is that all consistent and asymptotically normal adjusted estimators are in class (2.3) or asymptotically equivalent to the estimators in this class. So we do not lose generality by only studying this augmented method.

In practice, often $E(m(Y, A, \theta)|A = a, \mathbf{X}) = B(a, \theta)[E(Y|A = a, \mathbf{X}) - f(a, \theta)], a = 0, 1$. Then it is equivalent to developing a working model for $E(Y|\mathbf{X}, A = a)$. (Zhang et al., 2008)

For continuous Y and binary A using the practical strategy above, the class of consistent, asymptotic normal and robust estimators for θ could be expressed exactly or asymptotically as

$$\bar{Y}_1 - \bar{Y}_0 - \frac{1}{n} \sum_{i=1}^n \{(A_i - \hat{\pi})[(1 - \hat{\pi})^{-1}h_0(\mathbf{X}_i) + \hat{\pi}^{-1}h_1(\mathbf{X}_i)]\} = \hat{\mu}_1 - \hat{\mu}_0, \quad (2.5)$$

where $n_1 = \sum_{i=1}^n A_i$, $\hat{\pi} = n_1/n$, $\hat{\mu}_1 = \bar{Y}_1 - \frac{1}{n} \sum_{i=1}^n (A_i - \hat{\pi})\hat{\pi}^{-1}h_1(\mathbf{X}_i)$, $\hat{\mu}_0 = \bar{Y}_0 + \frac{1}{n} \sum_{i=1}^n (A_i - \hat{\pi})(1 - \hat{\pi})^{-1}h_0(\mathbf{X}_i)$. Note that $\bar{Y}_1 - \bar{Y}_0$ is the unadjusted estimator and the rest of the terms in (2.5) could be viewed as the augmentation term. $\hat{\mu}_1$ and $\hat{\mu}_0$ are consistent for $E(Y|A = 1)$ and $E(Y|A = 0)$ respectively and the augmentation term converge to 0 by randomization. From (2.4), the augmented estimator for the final analysis is denoted as

$$\hat{\theta}_{aug} = \bar{Y}_1 - \bar{Y}_0 - \frac{1}{n} \sum_{i=1}^n \{(A_i - \hat{\pi})[(1 - \hat{\pi})^{-1}q(\hat{\zeta}_0, \mathbf{g}(\mathbf{X}_i)) + \hat{\pi}^{-1}q(\hat{\zeta}_1, \mathbf{g}(\mathbf{X}_i))]\}, \quad (2.6)$$

where one postulates models $E(Y|A = a, \mathbf{X}) = q(\zeta_a, \mathbf{g}(\mathbf{X})) = \zeta_a^T \mathbf{g}(\mathbf{X})$, $a = 0, 1$.

However, the working regression model we build in the augmented method above requires full covariates information. When the baseline covariates are missing, our ability to make an accurate inference on treatment effect may be compromised and is concretely discussed in Section 2.4.

2.3.2 Methods for handling missing covariates

Here we review four general approaches for handling multivariate missing data in a regression setting.

1. Complete-case analysis: When baseline covariates are missing for certain subjects, it is simple to discard the incomplete subjects and analyze the data with complete information, which is referred to as complete-case analysis (CCA). However, CCA may lose efficiency due to the unnecessary deletion of sample size and lead to an inconsistent estimator when the data is not missing completely at random (MCAR). Even multivariate missing data with a small missing probability for each variable will lead to a significantly reduced sample size.
2. Mean and conditional mean imputation: Mean imputation, where the missing covariates are filled by their observed means; and conditional mean imputation (regression imputation), where the missing values are estimated by linear regression using information from observed data. (Little and Rubin, 2019)

3. Multiple imputation: Multiple imputation (MI) is a general statistical technique for handling missing data. Rubin (2018) Instead of imputing a single value for each missing covariate, multiple imputed datasets are generated from the predictive distribution. The estimates calculated from multiple imputed data sets are combined to yield a single inference using Rubin’s rules. (Rubin, 2018) Multiple imputation by chained equations (MICE) (Raghunathan et al., 2001; Van Buuren, 2007) has become popular nowadays based on a set of imputation models, one for each missing variable. White et al. White et al. (2011) provide good guidance for MICE and *mice* is a well-documented R package for practical implementation.
4. Instead of imputing the missing values, the missing observations are set to a fixed number, and a missing indicator R is added to the working model. In randomized clinical trials, the missing indicator method usually leads to a consistent estimator when the outcome is continuous mainly because covariates are balanced by randomization. (Groenwold et al., 2012)

Though there is a large volume of literature discussing the missing-covariates problem, they aim at making inferences on regression coefficients. However, in this paper, we focus on the inference of treatment effect regardless of the misspecification of working models. Therefore, the general methods for handling missing data could not be directly applied to the robust covariate adjustment framework discussed above. To the best of our knowledge, there is limited literature on missing covariates in the context of robust covariate adjustment. Therefore, to obtain consistent, efficient, and robust estimators in practical analysis, we believe that a modified robust covariate adjustment framework is needed when the baseline covariates are missing.

2.4 Method

We now focus on the inference of θ using the augmented method discussed in Section 2.3.1 when some covariates $\mathbf{X}$ might be missing, with the additional constraints that treatment assignment is independent of missingness and baseline variables jointly.

2.4.1 Complete case analysis

For i.i.d data $(Y_i, \mathbf{X}_i, A_i), i = 1, \dots, n$, the working regression model we build in the augmented method described in Section 2.3.1 requires full covariates information. When some covariates are missing, a common practice is the CCA. For subject i , let us define $R_i^* = 1$ if

$\sum_{j=1}^p R_{ij} = p$ and $R_i^* = 0$ otherwise. That is $R_i^* = 1$ if subject i is fully observed and $R_i^* = 0$ if there exists missingness in baseline covariates. From (2.5), the class of estimators for the final data analysis when some covariates are missing could be expressed as

$$\hat{\theta}_{augcc} = \bar{Y}_1^* - \bar{Y}_0^* - \frac{1}{n} \sum_{i=1}^n \{(A_i - \hat{\pi})R_i^*[(1 - \hat{\pi})^{-1}q(\hat{\zeta}_0, \mathbf{g}(\mathbf{X}_i)) + \hat{\pi}^{-1}q(\hat{\zeta}_1, \mathbf{g}(\mathbf{X}_i))]\}, \quad (2.7)$$

where $\bar{Y}_1^* = \frac{\sum_{i=1}^n A_i Y_i R_i^*}{\sum_{i=1}^n A_i R_i^*}$, $\bar{Y}_0^* = \frac{\sum_{i=1}^n (1-A_i) Y_i R_i^*}{\sum_{i=1}^n (1-A_i) R_i^*}$. However, CCA might lead to potential bias. By conditional expectation argument, $\bar{Y}_1^* - \bar{Y}_0^*$ converges in probability to $E(Y|A = 1, R^* = 1) - E(Y|A = 0, R^* = 1)$, but not necessarily to $E(Y|A = 1) - E(Y|A = 0)$, which may lead to inconsistency of $\hat{\theta}_{augcc}$.

2.4.2 Imputation method

The second approach to achieve our inference goal is to first impute the missing $\mathbf{X}$ and then apply the augmented method to the imputed data set. For mean and regression imputation, the class of estimators for θ could be expressed as

$$\hat{\theta}_{augimp} = \bar{Y}_1 - \bar{Y}_0 - \frac{1}{n} \sum_{i=1}^n \{(A_i - \hat{\pi})[(1 - \hat{\pi})^{-1}q(\hat{\zeta}_0, \mathbf{g}(\mathbf{X}_i^{\text{imp}})) + \hat{\pi}^{-1}q(\hat{\zeta}_1, \mathbf{g}(\mathbf{X}_i^{\text{imp}}))]\}, \quad (2.8)$$

where the $\mathbf{X}_i^{\text{imp}} = (X_{i1}^{\text{imp}}, \dots, X_{ip}^{\text{imp}})^T$ and for example, $X_{i1}^{\text{imp}} = R_{i1}X_{i1} + (1 - R_{i1})E(X_{i1}|X_{i2}, \dots, X_{ip})$ if X_1 is the only missingness. As for multiple imputation, we could use (2.8) to obtain a single estimate for each imputed data set and then combine estimates from multiple imputed data sets using Rubin's rules. (White et al., 2011)

When one imputes missing X 's by their sample means, $\hat{\theta}_{augimp}$ is consistent for θ because $\bar{Y}_1 - \bar{Y}_0$ converges to $E(Y|A = 1) - E(Y|A = 0)$ and the augmentation term converges to 0 by randomization (see Appendix A.2). However, when implementing regression imputation or multiple imputation, there remains a question of whether we should include outcome Y in the imputation model. Though White et al. (2011) suggested including both covariates and outcome in the imputation model to avoid bias, this conclusion comes from a regression setting where the main purpose is to study the relationship between covariates and outcome. This may not necessarily be optimal for our augmented method. Suppose we include the outcome in the imputation model, the working model in the augmentation term now becomes a function of $\mathbf{X}, \mathbf{R}, Y$. Then the augmentation term does not converge to 0, leading to an inconsistent estimator for θ (see Appendix A.2). Therefore, it is recommended

to use mean imputation or regression imputation with covariates only.

2.4.3 General augmented method

The idea of the approaches discussed in Section 2.4.1 and Section 2.4.2 is to first eliminate missingness and then apply the augmented method to the 'fully observed' data set. The other way is to borrow the idea from the missing indicator approach, recognizing the missingness as part of covariates information and directly applying the augmented method. This is reasonable since missingness could be viewed as prognostic baseline variables (variables that are correlated with outcome) and adjusting for those variables could improve precision and power. For example, in a randomized trial on the effect of certain treatments for depression, patients with severe depression would be unwilling to participate in the interview and have missing data. In addition, when using the augmented method, treating missingness as covariates allows all patients to be used for estimating the treatment effect.

Specifically, for continuous covariates, missing values are replaced by a fixed value, typically zero. Thus both modified covariates and observation indicators are included in the observed data. For missing categorical variables, missing values are grouped into a 'missing category' and therefore an additional group is created for this category (Jones, 1996). For simplicity, we assume data only contains continuous $\mathbf{X}$, our observed data which contains missing information is then defined as $(Y_i, \mathbf{X}_i^*, A_i), i = 1, \dots, n$, where new covariates $\mathbf{X}_i^* = (X_{i1}R_{i1}, \dots, X_{ip}R_{ip}, R_{i1}, \dots, R_{ip})$. We assume that data vectors are independent, identically distributed (i.i.d), and $A \perp\!\!\!\perp (\mathbf{X}, \mathbf{R})$. The second assumption is reasonable if baseline covariates missingness is not related to randomization in randomized clinical trials, for example, baseline covariates information is collected before randomization. The class of consistent and asymptotic normal estimators for final analysis could be expressed as

$$\hat{\theta}_{augr} = \frac{\sum_{i=1}^n A_i Y_i}{\sum_{i=1}^n A_i} - \frac{\sum_{i=1}^n (1 - A_i) Y_i}{\sum_{i=1}^n (1 - A_i)} - \frac{1}{n} \sum_{i=1}^n \{ (A_i - \hat{\pi}) [(1 - \hat{\pi})^{-1} q(\hat{\zeta}_0, \mathbf{g}(\mathbf{X}_i^*)) + \hat{\pi}^{-1} q(\hat{\zeta}_1, \mathbf{g}(\mathbf{X}_i^*))] \}, \quad (2.9)$$

which is the same as in Section 2.3.1 except for including missingness as an additional covariate.

2.4.3.1 Variable selection for working models

As mentioned earlier, the closer the working model is to the true $E(Y|\mathbf{X}, A = k)$, the greater the efficiency gain over the unadjusted estimator. To achieve this, we need to con-

sider not only linear terms but also quadratic terms, two-way interactions, and even more complex functional forms in the working model since the true model might contain them (see Appendix A.4). However, when there exists multivariate missingness, the number of covariates involved in working regression models becomes large. For example, when considering linear basis function, $g(\mathbf{X}^*)$ in working models is up to $2p$ -dimensional compared with p -dimensional for $g(\mathbf{X})$. Therefore, traditional OLS is not feasible and variable selection methods are considered for the general augmented method.

In comparison with the traditional stepwise procedures, such as forward or backward selection with *ad hoc* entering/moving criteria, modern variable selection methods, for example, ridge regression, LASSO regression (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005) are preferred in the working model to make a valid inference on θ with practical sample size. Here we utilize the LASSO procedure with cross-validation for constructing the working model when the dimension of covariates is not small.

2.4.3.2 Imputation as two-step augmented method

By carefully studying estimator (2.8) in Section 2.4.2, we recognize that the imputation method indirectly introduced R and their interactions with non-missing covariates into the working model. Since $q(\hat{\zeta}_{\mathbf{a}}, \mathbf{g}(\mathbf{X}_i^{\text{imp}})), a = 0, 1$ is a nested model within a general augmentation working model, the imputation method can be regarded as a two-step augmentation and aligns with the general augmented method (See Appendix A.3). However, this imputation method leads to a working model that does not converge to the optimal when $\mathbf{R}$ is correlated with the outcome, potentially leading to a loss of efficiency for θ (see Appendix A.4). Therefore, we propose a modified imputation method: first, perform imputation using other covariates but not the outcome, then apply the augmented method for the imputed data set with the inclusion of additional covariates $\mathbf{R}$. Efficiency can be improved by this modified imputation method.

2.5 Simulation Studies

We carried out simulation studies based on 1000 Monte Carlo data sets, focusing on evaluating the performance of the proposed methods. We considered two levels of missing patterns (univariate missingness and multiple missingness), two different missing mechanisms (MCAR and MNAR), and two types of missing probabilities (equally missing and unequally missing).

2.5.1 Univariate missingness

In the first simulation, we considered a simple setting where $X = (X_{1i}, X_{2i})^T$ and only X_1 was missing. We generated covariates $\mathbf{X} = (X_1, X_2)^T \sim N(0, \mathbf{D})$, with $D_{11} = D_{22} = 1$ and $D_{12} = 0.5$. A was Bernoulli with $P(A = 1) = 0.5$. The outcome Y was generated according to $Y = (3 + X_1 - 0.5X_2)A + (2 + X_1 + 0.3X_2)(1 - A) + \epsilon$, with $\epsilon \sim N(0, 1)$. Under MCAR, we set the observation indicator R for X_1 as Bernoulli with $P(R = 1) = 0.8$. Under MNAR, we generated R through $P(R = 1|X_1, X_2, \epsilon) = \text{logit}^{-1}(1.85 + X_1 + \epsilon) \approx 0.8$. Here, the true values of the parameters of interest were $(\mu_0, \mu_1, \theta) = (2, 3, 1)$. For each data set, with $n=800$, we estimated θ by the proposed methods in Section 2.4, where the working models for each $a = 1, 2$ in the augmentation term were developed by fitting OLS for $E(Y|\mathbf{X}, A = a) = q_a(\mathbf{X}, \zeta_a)$. The estimators we are evaluating include:

1. ‘un’: the standard unadjusted estimator defined in (2.2).
2. ‘cc-aug’: the complete case analysis using the augmented method defined in (2.7), where $q_a(\mathbf{X}, \zeta_a) = (1, X_1, X_2)^T \zeta_a$.
3. The augmented estimators after imputation defined in (2.8).
 - (a) ‘meanimp-aug’: use the observed means to impute the missing covariates, then apply the augmented method, where $q_a(\mathbf{X}, \zeta_a) = (1, X_1^{imp}, X_2)^T \zeta_a$.
 - (b) ‘conimp-aug’: include X_2 in the imputation model, $q_a(\mathbf{X}, \zeta_a) = (1, X_1^{imp}, X_2)^T \zeta_a$.
 - (c) ‘conimp-y’: include Y and X_2 in the imputation model, $q_a(\mathbf{X}, \zeta_a) = (1, X_1^{imp}, X_2)^T \zeta_a$.
 - (d) ‘conimp-aug (MI)’: use MI to create multiple imputed datasets, where X_2 is included in the imputation model, $q_a(\mathbf{X}, \zeta_a) = (1, X_1^{imp}, X_2)^T \zeta_a$. Combine estimates from multiple imputed data sets using Rubin’s rules. (White et al., 2011) The MI was implemented using the `mice` package in *R* with 10 imputed datasets and 5 iterations.
 - (e) ‘conimp-y-aug (MI)’: include Y and X_2 in the imputation model, with other processes the same as ‘conimp-aug (MI)’.
4. ‘main effect-aug’: the direct augmented method defined in (2.9), where $q_a(\mathbf{X}^*, \zeta_a) = (1, X_1 R, R, X_2)^T \zeta_g$.
5. ‘conimp-aug-r’: the imputation and direct augmentation combined method defined in (2.9), where we include X_2 in the imputation model, $q_g(\mathbf{X}, \zeta_g) = (1, X_1^{imp}, X_2, R)^T \zeta_g$.

Table 2.1: Simulation results for univariate missingness based on 1000 Monte Carlo datasets, with a sample size of 800.

Estimators	$\hat{\mu}_0$		$\hat{\mu}_1$		$\hat{\theta}$				
	Bias	MC SD	Bias	MC SD	Bias	MC SD	AVE. SE	Cov. prob.	Rel. eff.
<i>MCAR</i>									
un	0.001	0.075	-0.002	0.068	-0.002	0.101	0.102	0.956	1.00
cc-aug	0.000	0.069	-0.002	0.067	-0.002	0.085	0.085	0.953	1.41
meanimp-aug	0.000	0.065	-0.002	0.063	-0.001	0.083	0.082	0.949	1.50
conimp-aug	0.000	0.065	-0.001	0.062	-0.001	0.082	0.081	0.944	1.55
conimp-y-aug	0.039	0.063	-0.040	0.061	-0.079	0.076	0.073	0.800	0.86
conimp-aug (MI)	0.000	0.065	-0.001	0.062	-0.001	0.082	0.087	0.967	1.54
conimp-y-aug (MI)	0.031	0.063	-0.033	0.061	-0.065	0.077	0.080	0.883	1.02
main effect-aug	0.000	0.065	-0.002	0.063	-0.001	0.083	0.082	0.951	1.50
conimp-aug-r	0.000	0.064	-0.002	0.062	-0.001	0.081	0.081	0.946	1.55
meanimp-ANCOVAI	0.000	0.067	-0.001	0.065	-0.001	0.083	0.082	0.952	1.50
conimp-ANCOVAI	0.000	0.066	-0.001	0.064	-0.001	0.081	0.081	0.946	1.55
main effect-ANCOVAI	0.000	0.067	-0.001	0.064	-0.001	0.083	0.082	0.954	1.50
<i>MNAR</i>									
un	0.002	0.078	0.002	0.066	0.000	0.103	0.102	0.936	1.00
cc-aug	0.339	0.067	0.274	0.060	-0.065	0.079	0.081	0.874	1.01
meanimp-aug	0.000	0.068	0.002	0.063	0.002	0.089	0.087	0.944	1.36
conimp-aug	0.000	0.068	0.002	0.062	0.002	0.086	0.085	0.944	1.43
conimp-y-aug	0.050	0.063	-0.046	0.060	-0.096	0.077	0.074	0.746	0.71
conimp-aug (MI)	0.000	0.068	0.002	0.062	0.002	0.087	0.090	0.959	1.42
conimp-y-aug (MI)	0.032	0.065	-0.029	0.060	-0.061	0.079	0.081	0.895	1.08
main effect-aug	0.002	0.064	0.001	0.059	-0.001	0.075	0.076	0.956	1.88
conimp-aug-r	0.002	0.063	0.001	0.059	-0.001	0.075	0.076	0.955	1.91
meanimp-ANCOVAI	0.001	0.070	0.003	0.064	0.002	0.089	0.087	0.940	1.36
conimp-ANCOVAI	0.001	0.070	0.003	0.063	0.002	0.086	0.085	0.944	1.43
main effect-ANCOVAI	0.003	0.066	0.001	0.060	-0.002	0.075	0.082	0.955	1.88

Notes: Bias, mean of Monte Carlo bias; MC SD, the Monte Carlo standard deviation; AVE.SE, the average of standard error estimate; Cov. prob., the coverage probability of a 95% Wald confidence interval; Rel. eff., the Monte Carlo mean square error of unadjusted estimator divided by that of every other estimator.

6. ‘main effect/meanimp/conimp-ANCOVAI’: the ANCOVAI estimator as defined in Tsiatis et al. (2008). The corresponding ‘main effect/meanimp/conimp-ANCOVAI’ estimator was obtained by substituting X_1 with $(X_1 R, R)/X_1^{imp}/X_1^{imp}$ in the original ANCOVAI model.

Results are shown in Table 2.1. The ‘conimp-y-aug’ and ‘conimp-y-aug (MI)’ lead to biased estimates under both MCAR and MNAR. This is because ‘conimp-y-aug’ and ‘conimp-y-aug (MI)’ include the outcome in the imputation model, and therefore do not necessarily converge to $E(Y|A = 1) - E(Y|A = 0)$ (see Appendix A.2). The ‘cc-aug’ leads to a biased estimate under MNAR as discussed in Section 2.4.1. All other estimators show negligible bias and moderate efficiency gain. The ‘conimp-aug’ leads to a more efficient estimator compared with ‘meanimp-aug’ because the working model of ‘conimp-aug’ is closer to the optimal $E(Y|A = a, \mathbf{X}^*)$ (see Appendix A.3 and A.4). A key message is that ‘main effect-aug’ and

‘conimp-aug-r’ achieve the best efficiency gain, especially under MNAR, for the same reason. The very similar results between ANCOVAI methods and augmented methods (e.g., ‘conimp-aug’ and ‘conimp-ANCOVAI’) indicate the asymptotic equivalence of ANCOVA method and augmented method under certain conditions (see Appendix A.5), allowing analysts to obtain the ANCOVA estimators using the corresponding augmented methods.

Because of the bias and low coverage probability, we do not consider ‘cc-aug’, ‘conimp-y-aug’, or ‘conimp-y-aug (MI)’ in further simulation studies. We also exclude the ANCOVAI estimators because of the asymptotic equivalence with augmented methods. We assume the data are MNAR in all later simulations to justify that the MAR assumption is unnecessary.

2.5.2 Multivariate missingness

We considered multivariate missingness and two types of missingness probabilities in the second simulation to represent the practical analysis situation. We generated covariates $\mathbf{X} = (X_1, \dots, X_{10})^T$ such that $X_1, X_6 \sim N(0, 1)$, $\log(X_8) \sim N(0, 1)$, $X_2 = 0.1X_1 + 0.9U_1$, $X_3 = 0.4X_1 + 0.4X_2 + 0.2U_2$, $X_4 = 0.3X_2 + 0.7U_3$, $X_5 = 0.1X_1 + 0.1X_6 + 0.3X_8 + 0.5U_4$, $X_7 = 0.5X_6 + 0.5U_5$, $X_9 = 0.2X_1 + 0.2X_6 + 0.6U_6$, and $X_{10} = 0.5X_1 + 0.5U_7$, where $U_l \sim N(0, 1), l = 1, \dots, 7$. A was Bernoulli with $P(A = 1) = 0.5$. The outcome Y was generated according to $Y = (3 + 0.6X_1 + 0.2X_3 + 0.3X_7 + 0.2X_9)A + (2 + 0.6X_1 + 0.3X_2 - 0.4X_8 + 0.2X_9)(1 - A) + \epsilon$, with $\epsilon \sim N(0, 1)$. Covariates X_1 to X_5 were missing, and we generated corresponding observation indicators $R_j, j = 1, \dots, 5$ through logistic regressions. We considered two types of missingness probabilities. For equally missing, e.g., $P(R_m = 1) = P(R_n = 1) = 0.9, \forall m \neq n, m, n = 1, \dots, 5$, we set $P(R_1 = 1|\mathbf{X}, \epsilon) = 1 - \text{logit}^{-1}(-2.4 + 0.3X_1 + 0.8\epsilon)$, $P(R_2 = 1|\mathbf{X}, \epsilon) = 1 - \text{logit}^{-1}(-2.2 + 0.2X_2)$, $P(R_3 = 1|\mathbf{X}, \epsilon) = 1 - \text{logit}^{-1}(-2.4 + 0.6\epsilon)$, $P(R_4 = 1|\mathbf{X}, \epsilon)$ as Bernoulli with $P(R_4 = 1) = 0.9$ and $P(R_5 = 1|\mathbf{X}, \epsilon) = 1 - \text{logit}^{-1}(-2.2 + 0.2X_1 + 0.3X_7)$. For unequally missing, e.g., $P(R_1) = 0.8, P(R_m) = 0.95, m = 2, 3, 4, 5$, we set $P(R_1 = 1|\mathbf{X}, \epsilon) = 1 - \text{logit}^{-1}(-1.6 + 0.3X_1 + 0.8\epsilon)$, $P(R_2 = 1|\mathbf{X}, \epsilon) = 1 - \text{logit}^{-1}(-3 + 0.2X_2)$, $P(R_3 = 1|\mathbf{X}, \epsilon) = 1 - \text{logit}^{-1}(-3.1 + 0.6\epsilon)$, R_4 as Bernoulli with $P(R_4 = 1) = 0.95$ and $P(R_5 = 1|\mathbf{X}, \epsilon) = 1 - \text{logit}^{-1}(-3 + 0.2X_1 + 0.3X_7)$. The parameters of interest were $(\mu_0, \mu_1, \theta) = (1.341, 3, 1.659)$. For each data set, $n=800$, we considered ‘un’, ‘conimp-aug’, ‘conimp-aug-r’, ‘main effect-aug’, and the following additional estimators:

1. ‘conimp-aug-r (MI)’: In MI, for each missing covariate, we include all other covariates in the imputation model. For example, for missing X_1 , we include $X_2, \dots, X_{10}$ in the imputation model. After imputation, the working model is specified by $q_g(\mathbf{X}, \zeta_g) = (1, X_1^{imp}, \dots, X_5^{imp}, X_6, \dots, X_{10}, R_1, \dots, R_5)^T \zeta_g$. The MI was implemented using the *mice* package in *R* with 10 imputed datasets and 5 iterations. As for ‘conimp-aug-r’,

Table 2.2: Simulation results for multivariate missingness based on 1000 Monte Carlo datasets when data are MNAR, with a sample size of 800.

Estimators	$\hat{\mu}_0$		$\hat{\mu}_1$		$\hat{\theta}$				
	Bias	MC SD	Bias	MC SD	Bias	MC SD	AVE. SE	Cov. prob.	Rel. eff.
Equally missing									
un	0.002	0.076	0.003	0.063	0.001	0.098	0.099	0.953	1.00
conimp-aug	0.003	0.065	0.003	0.058	0.000	0.082	0.078	0.944	1.46
conimp-aug-r	0.003	0.064	0.003	0.058	0.000	0.080	0.075	0.938	1.53
conimp-aug-r (MI)	0.003	0.064	0.003	0.057	0.000	0.080	0.076	0.943	1.53
main effect-aug	0.003	0.065	0.003	0.058	0.000	0.080	0.076	0.939	1.51
aug-LASSO (20 variables)	0.003	0.065	0.003	0.057	0.000	0.080	0.076	0.938	1.50
aug-LASSO (28 variables)	0.003	0.065	0.002	0.057	0.000	0.080	0.076	0.936	1.50
aug-LASSO (115 variables)	0.004	0.065	0.002	0.058	-0.002	0.081	0.076	0.934	1.49
Unequally missing									
un	-0.004	0.074	-0.001	0.063	0.003	0.096	0.099	0.964	1.00
conimp-aug	-0.004	0.063	-0.002	0.056	0.002	0.077	0.079	0.951	1.55
conimp-aug-r	-0.004	0.063	-0.002	0.055	0.002	0.074	0.075	0.947	1.70
conimp-aug-r (MI)	-0.004	0.063	-0.002	0.055	0.002	0.074	0.075	0.951	1.69
main effect-aug	-0.004	0.063	-0.001	0.056	0.002	0.075	0.076	0.950	1.65
aug-LASSO (20 variables)	-0.004	0.063	-0.001	0.056	0.002	0.074	0.075	0.946	1.66
aug-LASSO (28 variables)	-0.004	0.063	-0.002	0.056	0.002	0.074	0.075	0.945	1.66
aug-LASSO (115 variables)	-0.002	0.064	-0.001	0.056	0.001	0.076	0.075	0.945	1.61

Notes: The ‘aug-LASSO (20 variables)’ considers only interactions between R_1 (the highest missing rate) and fully observed covariates. The ‘aug-LASSO (28 variables)’ considers only interactions between R_1 (the highest missing rate) and covariates defined in (2.9). The ‘aug-LASSO (115 variables)’ considers all interactions.

the conditional mean imputation was implemented using the `mice` package in *R* with 1 imputed dataset and 1 iteration.

2. ‘aug-LASSO’: the direct augmented estimator defined in (2.9), including main effects and two-way interaction terms in the working model. LASSO was used in the working model for variable selection. The LASSO procedure was implemented using `glmnet` package in *R* with 10-fold cross-validation based on mean squared error to choose the tuning parameter.

Results are shown in Table 2.2. All methods lead to unbiased estimators when data are MNAR. Compared with the unadjusted estimator, all other estimators yield considerable efficiency gains. MAR is not required to achieve consistency, efficiency, and robustness if robust covariate adjustment is carried out properly. It’s worth noting that ‘r-aug-lasso (115 variables)’, which considers all two-way interactions in the working model, should be closer to optimal $E(Y|A = a, \mathbf{X}^*)$. Therefore, theoretically, it should achieve better efficiency gains compared with other estimators. In contrast, ‘r-aug-lasso (115 variables)’ fails to include important covariates because of the finite sample size effect and hence might lead to considerable efficiency loss. MI performs well but is no better than single regression

imputation (e.g., ‘conimp-aug-r vs. ‘conimp-aug-r (MI)’). Again, the direct augmented method (‘main effect-aug’/‘aug-LASSO’) and the combined method (‘conimp-aug-r’) achieve better efficiency gains than the imputation method (‘conimp-aug’) in general.

2.5.3 Multivariate missingness with more covariates

To emphasize the better performance of the direct augmented method and the combined method over the imputation method, in the third simulation, we considered more missing covariates and data are MNAR. We generated covariates $\mathbf{X} = (X_1, \dots, X_{30})^T$ such that $X_1, X_4, X_6, X_{11}, X_{15}, X_{17}, X_{20}, X_{22}, X_{23}, X_{24}, X_{27}, X_{28}, X_{30} \sim N(0, 1)$, $\log(X_8) \sim N(0, 1)$, $X_{16} \sim N(1, 1)$, $X_9 \sim \text{Bernoulli}(0.3)$, $X_{12} \sim \text{Bernoulli}(0.7)$, $X_{18} \sim \text{Bernoulli}(0.4)$, $X_{26} \sim \text{Bernoulli}(0.1)$, $X_{29} \sim \text{Bernoulli}(0.5)$, $X_2 = 0.3X_1 + 0.3X_{11} + 0.4U_1$, $X_3 = 0.5X_{12} + 0.5U_2$, $X_5 = 0.1X_1 + 0.3X_4 + 0.4X_6 + 0.2U_3$, $X_7 = 0.3X_{17} + 0.3X_{20} + 0.2X_{22} + 0.2U_4$, $X_{10} = 0.5X_9 + 0.5U_5$, $X_{13} = 0.4X_1 + 0.4X_{11} + 0.2U_6$, $X_{14} = 0.2X_6 + 0.2X_{20} + 0.3X_{30} + 0.3U_7$, $X_{19} = 0.3X_{11} + 0.2X_{24} + 0.2X_{27} + 0.3U_8$, $X_{21} = 0.3X_6 + 0.4X_9 + 0.3U_9$ and $X_{25} = 0.2X_1 + 0.2X_4 + 0.2X_{12} + 0.2X_{15} + 0.2U_{10}$, where $U_l \sim N(0, 1), l = 1, \dots, 10$. A was Bernoulli with $P(A = 1) = 0.5$. The outcome Y was generated according to $Y = (3 + 0.6X_1 + 0.2X_3 + 0.3X_7 + 0.2X_9 + 0.5X_{12} + 0.7X_{19} + 0.2X_{21})A + (2 + 0.6X_1 + 0.3X_2 - 0.4X_8 + 0.2X_9 + 0.5X_{12} + 0.7X_{19} - 0.3X_{25})(1 - A) + \epsilon$, with $\epsilon \sim N(0, 1)$. Covariates X_1 to X_{10} were missing unequally, e.g., $P(R_1 = 0.8), P(R_m) = 0.95, m = 2, \dots, 10$. We generated the corresponding observation indicators through the following logistic regression for MNAR. $P(R_1 = 1) = 1 - \text{logit}^{-1}(-1.6 + 0.3X_1 - 0.8\epsilon)$, $P(R_2 = 1) = 1 - \text{logit}^{-1}(-3 + 0.2X_2)$, $P(R_3 = 1) = 1 - \text{logit}^{-1}(-2.7 + 0.6X_1 + 0.3X_2 - 0.4X_8 + 0.3\epsilon)$, R_4 as Bernoulli with $P(R_4 = 1) = 0.95$, $P(R_5 = 1) = 1 - \text{logit}^{-1}(-3 + 0.2X_1 + 0.3X_7)$, $P(R_6 = 1) = 1 - \text{logit}^{-1}(-3 + 0.4X_1 + 0.3X_7 - 0.2X_{20})$, R_7 as Bernoulli with $P(R_7 = 1) = 0.95$, $P(R_8 = 1) = 1 - \text{logit}^{-1}(-3 + 0.2X_{14} + 0.3X_{21})$, $P(R_9 = 1) = 1 - \text{logit}^{-1}(-3 + 0.6\epsilon)$ and $P(R_{10} = 1) = 1 - \text{logit}^{-1}(-3.7 + 0.6X_1 + 0.2X_3 + 0.3X_7 + 0.2X_9 + 0.5X_{12} + 0.7X_{19} + 0.2X_{21} + 0.3\epsilon)$. The parameters of interest were $(\mu_0, \mu_1, \theta) = (1.709, 3.504, 1.795)$. For each data set, $n=800$, we considered the same estimators as in Section 2.5.2. Results are shown in Table 2.3. Again, all methods lead to unbiased estimators, with maximum precision achieved by the direct augmented method (‘main effect-aug’/‘aug-LASSO’) and the combined method (‘conimp-aug-r’).

2.5.4 Simulation based on ACTG 175 trial

We then conduct the last simulation based on the ACTG 175 trial described in Section 2.6, replicating the real-world data. For each simulated data set, continuous baseline variables

Table 2.3: Simulation results for 30 missing covariates based on 1000 Monte Carlo datasets, with a sample size of 800.

Estimators	$\hat{\mu}_0$		$\hat{\mu}_1$		$\hat{\theta}$				
	Bias	MC SD	Bias	MC SD	Bias	MC SD	AVE. SE	Cov. prob.	Rel. eff.
un	-0.003	0.077	-0.001	0.066	0.002	0.099	0.100	0.958	1.00
conimp-aug	-0.002	0.067	-0.003	0.059	-0.001	0.079	0.077	0.946	1.57
conimp-aug-r	-0.002	0.067	-0.002	0.058	0.000	0.076	0.074	0.945	1.68
conimp-aug-r (MI)	-0.002	0.067	-0.002	0.058	0.000	0.076	0.075	0.945	1.67
main effect-aug	-0.002	0.067	-0.002	0.059	0.000	0.077	0.074	0.943	1.64
aug-LASSO (60 variables)	-0.002	0.067	-0.002	0.059	0.000	0.077	0.075	0.944	1.65
aug-LASSO (78 variables)	-0.002	0.067	-0.002	0.059	0.000	0.077	0.075	0.945	1.65

Notes: The ‘aug-LASSO (60 variables)’ considers only interactions between R_1 (the highest missing rate) and fully observed covariates. The ‘aug-LASSO (78 variables)’ considers only interactions between R_1 (the highest missing rate) and covariates defined in (2.9).

were generated from a multivariate normal distribution with empirical mean and variance calculated based on all 2139 patients. Binary baseline indicators and treatment A were generated from independent Bernoulli distributions using corresponding observed proportions. We generated $CD4$ count (cells/mm³, Y) at 20 ± 5 weeks based on:

$$\begin{aligned}
E(Y|A = 0, \mathbf{X}) \approx & -79.705 + 1.599(CD4) - 0.0007(CD4)^2 \\
& - 0.107(CD4 \times HEMO) - 0.005(CD4 \times WT) \\
& + 0.013(WT \times KARN) - 0.040(CD8 \times HIST) \\
& - 23.199(HEMO \times RACE)
\end{aligned}$$

$$\begin{aligned}
E(Y|A = 1, \mathbf{X}) \approx & 95.445 + 1.100(CD4) - 0.005(CD4)^2 - 142.288(HOMO) \\
& - 0.178(CD4 \times DRUG) - 0.087(CD4 \times RACE) + 0.033(CD8 \times HEMO) \\
& - 0.014(CD8 \times HOMO) - 0.021(CD8 \times HIST) - 0.720(AGE \times HIST) \\
& - 0.554(AGE \times SYMP) - 0.706(WT \times HEMO) + 1.282(WT \times DRUG) \\
& + 1.688(KARN \times HOMO) - 28.321(DRUG \times RACE) \\
& - 45.337(DRUG \times SEX) + 35.981(DRUG \times HIST) \\
& + 24.032(RACE \times HIST) - 3.602(SEX \times HIST)
\end{aligned}$$

with treatment specific variance $Var(Y|A = 0, \mathbf{X}) \approx (95.82)^2$ and $Var(Y|A = 1, \mathbf{X}) \approx (115.63)^2$. (Tsiatis et al., 2008)

Let $R_j, j = 1, \dots, 5$ be the corresponding response indicators for baseline $CD4$ count, drugs, Karnofsky score, antiretroviral history, and symptomatic status, respectively. Under MNAR, the observation indicators were generated through the following logistic regression.

For equally missing, e.g., $P(R_m = 1) = P(R_n = 1) = 0.9, \forall m \neq n, m, n = 1, \dots, 5$, we set $P(R_1 = 1|\mathbf{X}, \epsilon) = 1 - \text{logit}^{-1}(-3.6 + 0.003\text{CD4} + 0.008\epsilon)$, $P(R_2 = 1|\mathbf{X}, \epsilon) = 1 - \text{logit}^{-1}(-2.2 + 0.2\text{DRUG})$, $P(R_3 = 1|\mathbf{X}, \epsilon) = 1 - \text{logit}^{-1}(-2.4 + 0.006\epsilon)$, $P(R_4 = 1|\mathbf{X}, \epsilon)$ as Bernoulli with $P(R_4 = 1) = 0.9$ and $P(R_5 = 1|\mathbf{X}, \epsilon) = 1 - \text{logit}^{-1}(-3.2 + 0.002\text{CD4} + 0.3\text{SEX})$. For unequally missing, e.g. $P(R_1) = 0.8, P(R_m) = 0.95, m = 2, 3, 4, 5$, we set $P(R_1 = 1|\mathbf{X}, \epsilon) = 1 - \text{logit}^{-1}(-2.6 + 0.003\text{CD4} + 0.008\epsilon)$, $P(R_2 = 1|\mathbf{X}, \epsilon) = 1 - \text{logit}^{-1}(-2.9 + 0.2\text{DRUG})$, $P(R_3 = 1|\mathbf{X}, \epsilon) = 1 - \text{logit}^{-1}(-3.1 + 0.006\epsilon)$, $P(R_4 = 1|\mathbf{X}, \epsilon)$ as Bernoulli with $P(R_4 = 1) = 0.95$ and $P(R_5 = 1|\mathbf{X}, \epsilon) = 1 - \text{logit}^{-1}(-3.9 + 0.002\text{CD4} + 0.3\text{SEX})$. The parameters of interest were $(\mu_0, \mu_1, \theta) = (315.483, 378.458, 62.975)$. For each data set, $n=2000$, we considered ‘un’, ‘main effect-aug’, ‘conimp-aug-r’, and ‘aug-LASSO’. Simulation results are shown in Table 2.4. All estimators are unbiased, with similar precision gains achieved by the direct augmented method and the combined method.

Table 2.4: Simulaton results based on the ACTG data with 1000 Monte Carlo datasets and a sample size of 2000.

Estimators	$\hat{\mu}_0$		$\hat{\mu}_1$		$\hat{\theta}$				
	Bias	MC SD	Bias	MC SD	Bias	MC SD	AVE. SE	Cov. prob.	Rel. eff.
Equally missing									
un	0.175	5.835	0.073	3.856	-0.102	6.901	6.985	0.955	1.00
conimp-aug-r	0.600	4.785	-0.149	3.688	-0.749	5.247	5.254	0.946	1.70
main effect-aug	0.646	4.817	-0.157	3.681	-0.804	5.268	5.249	0.949	1.68
aug-LASSO (158 variables)	0.552	4.825	-0.136	3.680	-0.688	5.272	5.188	0.949	1.69
Unequally missing									
un	-0.371	5.959	-0.018	3.767	0.353	7.149	6.980	0.946	1.00
conimp-aug-r	0.330	4.953	-0.323	3.530	-0.652	5.437	5.329	0.939	1.71
main effect-aug	0.402	4.956	-0.337	3.530	-0.738	5.438	5.330	0.936	1.70
aug-LASSO (37 variables)	0.357	4.956	-0.324	3.520	-0.681	5.428	5.274	0.931	1.71

Notes: The ‘aug-LASSO (37 variables)’ considers only quadratic terms of continuous variables and interactions between R_1 (the highest missing rate) and defined in (2.9). The ‘aug-LASSO (158 variables)’ considers all two-way interactions and quadratic terms of continuous variables.

2.6 Application

We demonstrated the proposed methods by application to data from AIDS Clinical Trials Group protocol 175 (ACTG), with the aim of estimating marginal treatment effect on the $CD4$ count (cells/mm³, Y) at 20 ± 5 weeks post-baseline. This trial was randomized to four equal proportions (Hammer et al., 1996) and we consider the following two groups. One group is zidovudine (ZDV) monotherapy ($A = 0, n_0 = 532$), and the other group consists of three treatment arms combined (ZDV+ didanosine (ddI), ZDV +zalcitabine and ddI monotherapy) ($A = 1, n_1 = 1607$). Baseline covariates X used in augmentation are listed in

Table 2.5: Data structure for ACTG data

Covariates	Variable Type	Measurement	Summary Statistics
Age	Continuous	Years	35.25 (8.71)
Weight	Continuous	Kg	75.13 (13.26)
Baseline CD4 count	Continuous	Cells/mm ³	350.50 (118.57)
Baseline CD8 count	Continuous	Cells/mm ³	986.63 (480.20)
Karnofsky score	Continuous	Scale of 0-100	95.45 (5.90)
Gender	Binary	0 = Female, 1 = Male	0.83 (1771)
Race	Binary	0 = White, 1 = Nonwhite	0.29 (617)
Hemophilia	Binary	0 = No, 1 = Yes	0.084 (180)
Homosexual activity	Binary	0 = No, 1 = Yes	0.66 (1414)
History of intravenous drug use	Binary	0 = No, 1 = Yes	0.13 (281)
Antiretroviral history	Binary	0 = Naive, 1 = Experienced	0.59 (1253)
Symptomatic status	Binary	0 = Asymptomatic	0.17 (370)

Notes: For summary statistics, mean (standard deviation) is presented for continuous variables, while proportions (counts) are provided for binary variables.

Table 2.5.

All covariates are observed in the ACTG data. To mimic the missing data problem frequently encountered in practice, we manually created two missing scenarios (equally missing and unequally missing) under MNAR based on logistic regressions specified in Section 2.5.4. Specifically, baseline *CD4* count is MNAR, and the missingness depends on itself and the error term; drugs is MNAR, and the missingness depends on itself; Karnofsky score is MNAR, and the missingness depends on the error term; antiretroviral history is MCAR; symptomatic status is MNAR, and the missingness depends on the missing *CD4* count and the error term. Note that the error term in logistic regressions specified in Section 2.5.4 is now substituted by residuals calculated based on the data.

The results are presented in Table 2.6. The relative efficiency in this table is computed as the ratio of the variance of the unadjusted estimate to those of the augmented ones. All methods indicate a significant treatment difference. All estimates of θ are very similar except for the unadjusted one. All augmented methods gain approximate 30% efficiency over the unadjusted one.

2.7 Discussion

In theory, we recognize that missing covariates do not pose additional challenges as long as missingness is independent of treatment assignment. One can directly apply the robust covariate adjustment framework by treating the missingness as an additional covariate. However, in practice, as the number of covariates and the amount of missingness increase, the working model becomes complex when applying robust covariate adjustment. To address this, we have proposed three different implementation strategies to handle the increased

Table 2.6: Estimates of β for the ACTG data, relative efficiency is calculated by standard error only

Estimators	$\hat{\mu}_0$	$\hat{\mu}_1$	$\hat{\theta}$	SE	Rel. eff.
Equally missing					
un	336.139	382.950	46.811	6.760	1.00
conimp-aug-r-LASSO	335.350	383.120	47.770	5.089	1.33
main effect-aug-LASSO	335.495	382.998	47.503	5.079	1.33
r-aug-LASSO (158 variables)	334.622	383.051	48.429	4.961	1.36
Unequally missing					
un	336.139	382.950	46.811	6.760	1.00
conimp-aug-r	334.856	383.140	48.284	5.201	1.30
main effect-aug	335.049	383.149	48.099	5.232	1.29
aug-LASSO (158 variables)	335.511	383.287	47.777	5.119	1.32

dimensionality in working models caused by missingness:

- Consider the basis function of $\mathbf{X}^*$ and use the variable selection method for working models.
- First impute the missing covariates using only covariates and then apply the augmented method to the imputed data set.
- First perform the imputation using only covariates, then apply the augmented method for the imputed data set with the inclusion of the additional covariates $\mathbf{R}$.

We realized that all these strategies can be thought of as augmented methods with different ways to build the working models and handle the high dimensional sets of ‘covariates’ $\mathbf{X}^*$.

Based on the simulation results, the first and the third strategies are recommended. We also recognized that knowledge learned in missing data literature on imputation does not necessarily apply in this setting. Missing at random assumption is not required. The imputation model should not include the outcome. Multiple imputation has good performance but is not better than regression imputation.

CHAPTER 3

Cautious Notes of Multiple Imputation in Survival Analysis: a Simulation Study

3.1 Introduction

Missing covariates are a very common problem in clinical and epidemiologic studies and may lead to biased estimates if we do not account for missingness. In survival analysis, the missing covariates problem becomes more challenging since the outcome is subject to censoring.

Without censoring, statistical methods have been extensively studied for handling missing covariates (Rubin, 1976; Little, 1992; Schafer and Graham, 2002; White and Carlin, 2010; Little and Rubin, 2019). One simple approach is the complete case (CC) analysis which uses cases with all fully observed covariates. However, CC analysis may lose efficiency due to the unnecessary deletion of sample size and yield potential bias when missingness depends on the outcome (Little, 1992). To correct bias and improve the efficiency of CC, several approaches have been proposed. These approaches generally require further modeling assumptions, either regarding the model for missing/selection probability, or with respect to the conditional model of covariates subject to missingness given the observed data, or both. The inverse probability weighted (IPW) estimators of (Robins et al., 1994) uses the inverse of selection probability as the weight and is consistent when data are missing at random (MAR i.e., the probability of missingness depends only on the observed data). However, it may be inefficient if the missing rate is high (Wang and Chen, 2001). By assuming an additional model for missing covariates and adding an augment term to IPW, Lipsitz et al. (1999) further proposed the augmented inverse probability weighted (AIPW) estimator for missing covariates. The AIPW estimator has a double robustness property and is more efficient compared with IPW estimator. On the other hand, imputation methods, including single and multiple imputation methods, poses a popular alternative to CC since it is relative simple to implement (Little and Rubin, 2019). The missing covariates are predicted from

the model involving the missing covariates and then are filled in, leading to the complete data and following standard analysis. Single imputation methods impute one value for each missing covariates (e.g. mean imputation), however, the standard error of the coefficients obtained by complete data is too small. To account for imputation uncertainty, Rubin (1978, 1987) proposed multiple imputation (MI). The main idea of this approach is to first generate multiply complete data sets by imputing missing values, then analyze these imputed data sets with standard complete data analysis procedures and combine the multiple analysis results to yield a single inference using the Rubin’s rules (Rubin, 1987).

With censoring outcome in survival analysis, several approaches have been proposed for handling missing covariates. Chen and Little (1999) considered nonparametric maximum likelihood to estimate regression parameters in Cox regression with missing covariates, while Wang and Chen (2001) extended the general AIPW estimation framework to solve this problem. Another alternates is MI. Both in survival or in general setting, MI has become an extremely popular approach for dealing with missing covariates in practice since it is more efficient than CC and relatively simple to implement by existing software (Van Buuren et al., 1999).

In survival analysis, most of the approaches mentioned above require data are MAR (i.e. the probability of missingness depends on the observed time, censoring indicator and fully observed covariates). Several literature have investigated the performance of MI through simulation studies under MAR (White and Royston, 2009; Qi et al., 2010; Marshall et al., 2010; Hsu and Yu, 2019). However, only a limited number of studies considered the robustness of MI under missing not at random (MNAR) assumption and imposed missingness on two more covariates (Marshall et al., 2010). Few literature (White and Royston, 2009; Marshall et al., 2010) failed to demonstrate the potential bias of CC under MAR except Qi et al. (2010) and Hsu and Yu (2019). They overlooked simulation scenarios where missingness may depend on survival outcomes and therefore their simulation scenarios were not comprehensive. Furthermore, by considering the ways in which MAR arise in survival, Rathouz (2007) examined the MAR assumption and argued that MAR is hard to justify in practice when missingness depends on outcome. In contrast to MAR, Rathouz (2007) proposed two interpretable missisng data assumptions: failure ignorable MAR (i.e. missingness does not depend on failure time) and censoring-ignorable MAR (i.e. missingness does not depend on censoring time but may depend on failure time). There is a lack of relevant simulation study assessing the performance of MI when considering the plausibility of MAR in survival analysis.

In this article, we focus on settings in which the interests lies in fitting a Cox model in survival analysis, with multiple potentially missing covariates. By varying the amount of

missingness and missing mechanisms, this paper discuss the MAR plausibility and explore the performance of MI and other frequently used methods in practice through an extensive simulation study. The aim of this paper is to summarize the results under different scenarios and provide a practical guidance for handling missing covariates in survival analysis.

In Section 3.2, we introduce the basic notations and review commonly used methods for handling missing covariates. In Section 3.3, we discuss the plausibility of MAR in survival analysis, introduce FIMAR and CIMAR and summarize different missing mechanisms used for comparison. In Section 3.4, we report the results of simulation. A discussion follows in Section 3.5.

3.2 Notations and missing data method

3.2.1 Notations

We assume that $(U_i, \Delta_i, \mathbf{X}_i, \mathbf{Z}_i), i = 1, \dots, n$ are independent and identically distributed. For a certain subject i , let T_i denote the failure time and C_i denote the censoring time. Let $U_i = \min(T_i, C_i)$ be the observed time and $\Delta_i = I(T_i \leq C_i)$ be the event indicator. Let $\mathbf{X}_i$ be a p -dimensional covariate vector that may be missing, and $\mathbf{Z}_i$ be a q -dimensional vector with fully observed covariates. Let $\mathbf{R}_i$ be the covariate observation indicator vector, a p -dimensional vector with components $R_{ij} = 1$ if X_{ij} is observed and $R_{ij} = 0$ if X_{ij} is missing. Let $R_i^* = 1$ if X_i is fully observed and $R_i^* = 0$ otherwise. Lastly, we assume T has a hazard function of $\lambda(t; \mathbf{X}_i, \mathbf{Z}_i) = \lambda_0(t) \exp(\beta_{\mathbf{X}}^T \mathbf{X}_i + \beta_{\mathbf{Z}}^T \mathbf{Z}_i)$, where $\lambda_0(t)$ is an unspecified baseline hazard function and $\beta = (\beta_{\mathbf{X}}^T, \beta_{\mathbf{Z}}^T)^T$ are parameters of interest.

3.2.2 Missing data methods in survival analysis

In this section, we review methods that are popular and simple to implement in practice for dealing with missing covariates in survival setting described as follows.

1. Complete case analysis: When baseline covariates are missing, it is simple to discard the incomplete subjects and analyze the data with complete information. Specifically, CC analysis of regression coefficients based on data $\{U_i, \Delta_i, \mathbf{X}_i, \mathbf{Z}_i : R_i^* = 1, i = 1, \dots, n\}$. Though it is easy to implement, CC analysis may loss efficiency due to the unnecessary deletion of sample size and lead to inconsistent estimator when the data is not missing completely at random (MCAR).
2. Single imputation: In order to prevent the loss of information contained in the incomplete subjects, an alternative way is to use the observed subjects to impute the

missing covariates. Missing covariates could be replaced by their unconditional sample means or a predicted value conditional on the observed data (Little and Rubin, 2019). For example, if X_1 is the only missingness. We could impute missing X_1 by $E(X_1|Z_1, \dots, Z_q, U, \Delta) = \gamma_0 + \sum_{j=1}^q \gamma_j X_j + \gamma_{q+1} U + \gamma_{q+2} \Delta$, where the γ s are estimated using complete cases (Marshall et al., 2010). Though single imputation (SI) maintains the sample size, it may lead to potential bias even assuming MCAR. This method also have the problem of 'too small' standard error, leading to an invalid inference.

3. Multiple imputation: MI as become an extremely popular approach for dealing with missing data in practice. We searched PubMed for any study using MI in Cox model from January 1st, 2010, using the terms ((multiple imputation) AND ((survival) OR (Cox)) AND ((missing data) OR (missing covariates))) in all fields. We found that 657 papers are using MI for missing data in survival analysis.

Instead of imputing a single value for each missing covariates, multiple imputed datasets are generated from the predictive distribution. There are two general approaches in MI: joint modeling (JM) (Schafer, 1997) and multiple imputation by chained equations (MICE) (Van Buuren et al., 1999). We mainly focus on MICE in this article since Cox model is semiparametric while JM involves specifying a specific multivariate distribution for data (Qi et al., 2010). The details of the standard MICE algorithm are described below (Van Buuren and Groothuis-Oudshoorn, 2011). For each univariate missing covariate, MICE initializes the missing values by drawing random samples with replacement from observed values. The appropriate imputation model is specified according to the type of each missing covariate (e.g. linear regression model for a continuous variable and logistic model for a binary variable). Then imputation is done in a variable-by-variable fashion and each imputation model conditions on the fully observed variables and the most recent imputed values of other missing covariates. One iteration consists of cycling through all missing covariates. To stabilize the result, several iterations are repeated until converge to generate one imputed data set and M multiple imputed data sets are generated by repeating the MICE algorithm M times.

Each of M imputed datasets is then analyzed by standard method, for example, Cox regression model. The estimates calculated from multiple imputed data sets are combined by Rubin's rules to obtain the overall estimates and variance (Rubin, 1987). Suppose $\hat{\beta}_{\mathbf{k}}$ is the estimate from k th imputed data set and $\mathbf{W}_{\mathbf{k}}$ is the corresponding estimated covariance variance matrix. The combined estimate $\hat{\beta} = \frac{1}{M} \sum_{k=1}^M \hat{\beta}_{\mathbf{k}}$ and the overall

Table 3.1: Imputation methods

Method label	Method description
CC	Complete case analysis
SI mean	Single imputation: Mean imputation
SI norm pred	Single imputation: using predicted value of linear regression
SI pmm	Single imputation: bayesian imputation with predictive mean matching
MICE pmm	MICE: bayesian imputation with predictive mean matching
MICE pmm nelsonaalen	MICE: bayesian imputation with predictive mean matching

$$\text{variance } Var(\hat{\beta}) = W + (1 + \frac{1}{M})\mathbf{B} \text{ where } \mathbf{W} = \frac{1}{M} \sum_{k=1}^M \mathbf{W}_k \text{ and } \mathbf{B} = \frac{1}{M-1} \sum_{k=1}^M (\hat{\beta}_k - \hat{\beta})^2.$$

Appropriate imputation models for missing covariates are important because, otherwise, it will bias the association between the covariates and outcomes. It is recommended that the outcome should be included in the imputation model (Moons et al., 2006). In survival analysis, finding the appropriate imputation model becomes more challenging due to censoring. An influential paper recommended that each incomplete covariate should be predicted from a imputation model including $U, \log U, \Delta$ and all other covariates (Van Buuren et al., 1999). White and Royston (2009) compared different imputation models for missing covariates and concluded that the appropriate imputation model should include Δ , the cumulative baseline hazard and other covariates. However, White and Royston (2009) discussed only the univariate missingness and the cumulative baseline hazard need to be approximated by the Nelson–Aalen estimator since the cumulative baseline hazard is unknown in practice. Qi et al. (2010) compared two popular imputation algorithms in MICE: Bayesian regression imputation and predictive mean matching. Qi et al. (2010) concluded that MICE with predictive mean matching has a slightly better performance.

Sophisticated methods such as AIPW and likelihood based methods have been carefully studied for dealing with missing covariates in survival analysis (Chen and Little, 1999; Wang and Chen, 2001; Qi et al., 2010), however, they are seldom used in practice. In this paper, we focus on methods that are frequently used in practice. CC analysis, three single imputation methods and two MICE techniques were investigated and summarised in Table 3.1. For all imputation approaches, the imputation model included $U, \log U, \Delta$ and all other covariates (Van Buuren et al., 1999) except for *MICE_pmm_nelsenallen*, where the event indicator Δ and the Nelson–Aalen estimator of $H(U)$ were included in the imputation model.

3.3 Missing data mechanisms

Rubin (1976) classified missing data problems into three categories. Data are said to be missing completely at random (MCAR) if the probability of missing is unrelated with data. A less rigid assumption would be missing at random (MAR), for which missing may depend on fully observed outcomes and covariates. If neither MCAR nor MAR holds, the mechanism is referred as missing not at random (MNAR).

Without censoring, the general MAR assumption about missing X is defined as

$$R \perp\!\!\!\perp X | (Y, Z), \quad (3.1)$$

where Y denotes the response variable in a regression model and Z denotes the fully observed covariates. In survival analysis problems with censoring, the MAR assumption is that the missingness of X is explicitly depend on fully observed outcomes U, Δ and covariates Z . This assumption,

$$R \perp\!\!\!\perp X | (U, \Delta, Z), \quad (3.2)$$

is generally termed MAR in survival analysis. Several approaches have been proposed to deal with missing covariates in Cox regression. The augmented inverse probability weighted (AIPW) method (Robins et al., 1994; Wang and Chen, 2001) uses two fully specified parametric models (given the observed data, one is the missing status model and the other is the missing covariates model) to account for missing covariates. The resulting estimators of regression coefficients of Cox model are doubly robust. Qi et al. (2010) further proposed fully augmented weighted estimators (FAWEs) in which using two non-parametric regression models to estimate the missingness and the missing covariates. However, the AIPW method is sensitive to model misspecification and FAWEs does not perform well when the dimension of observed data increases. MI (Rubin, 2004) is another widely used method in practice to deal with missing covariates in Cox regression. When models (imputation model and analysis model) are correctly specified, MI produce consistent and efficient estimates (White et al., 2011). We found that 285 papers are using MI for data imputation in Cox regression in recent 10 years. A further modification of MICE was proposed by Bartlett et al. (2015) when considering incompatibility between the imputation model and analysis model. Incompatible is defined if there exists no joint model satisfy these two models. For example, the default imputation model in MICE is linear, which may not be compatible with the Cox model in analysis stage. However, this method needs the analysis model is corrected specified.

Most of the methods described above require the data satisfied MAR assumption. If

missingness only depends on fully observed covariates, both (3.1) and (3.2) are well defined. However, when missingness of X is thought to depend on outcomes, (3.1) is well defined while (3.2) may not be plausible. Because when conditioning on fully observed covariates Z , the dependence of missingness on outcomes usually through some unmeasured intermediate variables since the missingness of covariates measured at baseline cannot depend on a future outcome directly. For example, a clinical outcome may be explained by income. Patients who have a higher education level may not be willing to report their incomes. But the education level is not measured in this study and may also reflect the clinical outcome. In this case, the presence or absence of income will predict the clinical outcome. In other settings, the missingness may be related with future failure time (Rathouz, 2007). Suppose a certain covariate may be measured only via a test but the test may not be safe for patients with poor prognosis. In this case, the presence or absence of the test predicts the future failure times through unmeasured covariate which indicate the poor prognosis. For MAR without censoring, we may consider the missingness of X depend on ϵ in a linear regression model $Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$, where the β_s are the parameters of interest. Therefore, the missingness of X depend on ϵ will lead to (3.1) and it is well defined. In survival setting, one obvious way to achieve MAR is when missingness is by design and explicitly dependent on (U, Δ) . When missingness depend on outcomes is by chance, we may consider missingness of X depend on failure time T or censoring time C though intermediate variables U_1 or U_2 respectively. However, the missingness of X depend on U_1 or U_2 will lead to $R \perp\!\!\!\perp X|(T, C, Z)$ but does not necessarily lead to (3.2), thus it may be plausible.

Rathouz (2007) critically explored various ways to obtain MAR through functions of T and C and proposed two alternative and interpretable missing data assumptions: censoring-ignorable missing at random (CIMAR) and failure-ignorable missing at random (FIMAR). To capture the situation where the missingness is related to failure time, but not to censoring, CIMAR is defined as

$$R \perp\!\!\!\perp (C, X)|(T, Z). \quad (3.3)$$

On the other hand, FIMAR allows missingness to depend on censoring time, but not on failure time, and is defined as

$$R \perp\!\!\!\perp (T, X)|(C, Z). \quad (3.4)$$

When additionally assumptions $C \perp\!\!\!\perp T|(X, Z)$ and $C \perp\!\!\!\perp X|Z$ are made, Rathouz (2007) concluded that MAR is difficult to justify in practice. One exception would that if one assumes FIMAR along with $C \perp\!\!\!\perp X|Z$.

In survival analysis, several simulation studies have considered the effects of different methods (e.g. AIPW, MI), but these have primarily restricted the missing data mecha-

Table 3.2: Missing mechanisms.

Abbreviations	Descriptions
MCAR	Missing complete at random
MAR general	Missing at random, assumption (3.2)
MAR FIMAR	Missing at random, assumption (3.4) with $C \perp\!\!\!\perp X Z$
MNAR XZ	Missing not at random, R depend on X and Z
MNAR FIMAR	Missing not at random, assumption (3.4) with $C \not\perp\!\!\!\perp X Z$
MNAR CIMAR	Missing not at random, assumption (3.3)

nisms under MAR and at most two missing covariates (White and Royston, 2009; Qi et al., 2010; Bartlett et al., 2015; Hsu and Yu, 2019). Few simulation studies considered both MAR and MNAR (Marshall et al., 2010), while it fail to demonstrate that CC may lead to potential bias when MAR and therefore the performance of MI over CC. The reason is they ignore a less well cited result, that is, when missingness depends on covariates only (observed or not), CC analysis yield consistent estimates (Little, 1992). More specifically, CC analysis focus on estimating the parameters based on the density of complete data $f(Y|R^* = 1, X, Z, \beta)$, which is not equivalent to the density of our interest $f(Y|X, Z, \beta)$. However, if $P(R^* = r|X, Z, Y) = P(R^* = r|X, Z)$ then $f(Y|R^* = 1, X, Z, \beta) = f(Y|X, Z, \beta)$ and this result has a straightforward extension to Cox model. With missing baseline covariates in survival analysis, the validity of CC does not fit the missing data mechanisms (MCAR/MAR/MNAR) exactly. When missingness of baseline covariates is independent of (T, C) , for example, if the data follow the MAR mechanism $f(R^*|U, X, Z) = f(R^*|Z)$ or even the MNAR mechanism $f(R^*|U, X, Z) = f(R^*|X, Z)$, CC produces valid inference. But not if the missingness depend on survival outcome. In addition, to the best of our knowledge, there remains a lack of simulation study comparing the existing methods under CIMAR or FIMAR for handling missing covariates in survival setting. Therefore, to capture all the missing covariates mechanism in survival analysis, it is worthwhile to consider not only the general MAR, but also CIMAR, FIMAR and missing depend on covariates only considering practical missingness generation procedure. In this paper, the missing mechanisms investigated were MCAR, two MAR mechanisms and three MNAR mechanisms which are summarised in table 1.

3.4 Simulation

We carried out simulation studies based on 1000 Monte Carlo data sets with sample size 1000, comparing the performance of the imputation methods described above based on Percentage bias (PB), i.e., monte Carlo bias of the estimated coefficients divided by the true value of

the parameter. A bias is considered large if the absolute value of PB is larger than 10, Monte Carlo mean square error (MC MSE) and Coverage probability (Cov. Prob.). For implementation of the multiple imputation, missing data were imputed $m = 5$ times and five iterations were set to impute each data set. The analysis model was a Cox PH model on X and Z .

We studied the performance of the imputation methods under different missing mechanisms summarised in Table 3.2. We considered both a single missing covariate and multiple missing covariates and varied the overall rate of missingness of 20%, 40% and 70%.

3.4.1 Simulation 1: two covariates, univariate missingness

We first considered the simple case with a single missing X and a single complete Z . The covariates X and Z were generated from a standard bivariate normal distribution with correlation $\rho = 0.5$. Survival times were simulated with hazard function $h(t|X, Z) = 0.07 \exp(X + 0.5Z)$. Censoring time were generated from exponential distribution with hazard 0.03 to give 35% censoring.

Under MCAR, We create missingness of X as a Bernoulli random variable, resulting in 20%, 40% and 70% missingness. Under MAR and MNAR, we create missingness of X from logistic regression model, i.e.

$$\text{logit}(P(R = 1)|Z) = \alpha_0 + \alpha_1 Z \quad (3.5)$$

when missingness depend on Z . The overall missing rate was determined by different choices of $\alpha = c(\alpha_0, \alpha_1)$. One special case would be that, in MNAR FIMAR, the censoring time were simulated using exponential distribution depending on missing X . Details of missing mechanisms can be found in Appendix.

PB and MSE of the missing X_1 are in Figure 3.1. CC is unbiased when the data are MCAR, MAR FIMAR, MNAR XZ, or MNAR FIMAR. Because CC is valid when missingness depend only on covariates (Little and Rubin, 2019). Other imputation methods are biased. Because we assume the nonlinear relationship between covariates and the survival outcomes while the imputation model for missing X is specified as linear. Therefore, the imputation model is misspecified and incompatible with the analysis model, resulting in biased estimates. Two MICE methods show small bias across all missing data mechanisms when the missing rate isn't too high (i.e., 70%). SI norm pred and SI mean have large bias with most missing data mechanisms and missing rate. Because of its poor properties, we do not discuss SI norm pred and SI mean in further simulation studies. As for MSE, CC performs the best for the most scenarios since it has the smallest bias and the MC SD is not that large compared

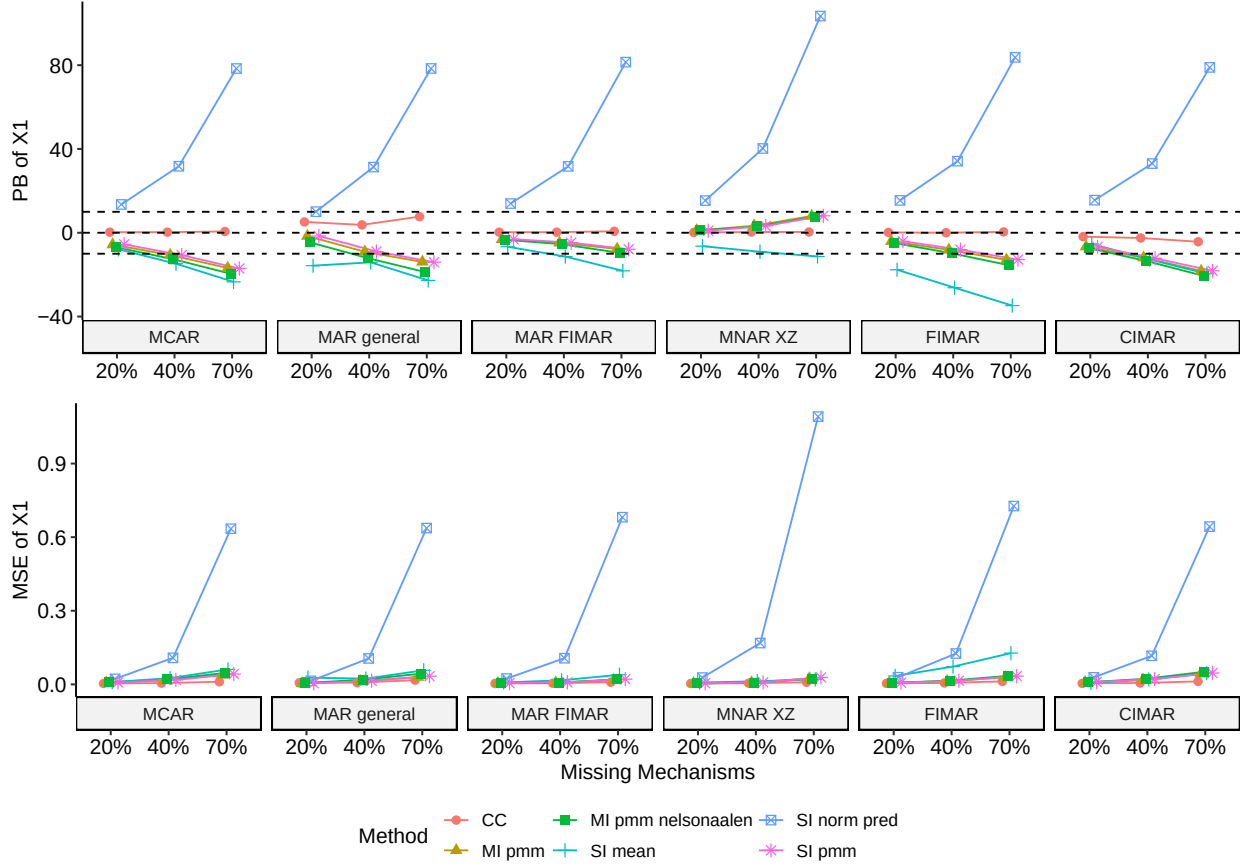


Figure 3.1: PB and MC MSE of X_1 in simulation 1. Dashed lines represent the $\pm 10\%$ PB.

with other methods across all missing mechanisms and missing rate. One exception is when missing rate is low (i.e., 20%) and data are MAR general. Two MI techniques outperform CC in this case. Z_1 follows a similar pattern for PB and MSE (see Appendix).

MC SD and coverage probability of the missing X_1 are in Figure 3.2. SI mean has the smallest SD CC remains around the nominal 95% level for all missing mechanisms and missing rate. of missingness

When X is missing complete at random (MCAR), CC is unbiased (Little and Rubin, 2019) while other imputation methods are biased. Because we assume the nonlinear relationship between covariates and the survival outcomes while the imputation model for missing X is specified as linear. Therefore, the imputation model is misspecified and incompatible with the analysis model, resulting in biased estimates. In terms of efficiency, MICE pmm and MICE pmm nelsenallen slightly improved the efficiency compared with CC since the sample size is smaller when using CC. The standard error are 'too small' for all single imputation methods since single imputation does not account for uncertainty. Because of its poor properties, we do not report single imputation results in further simulations.

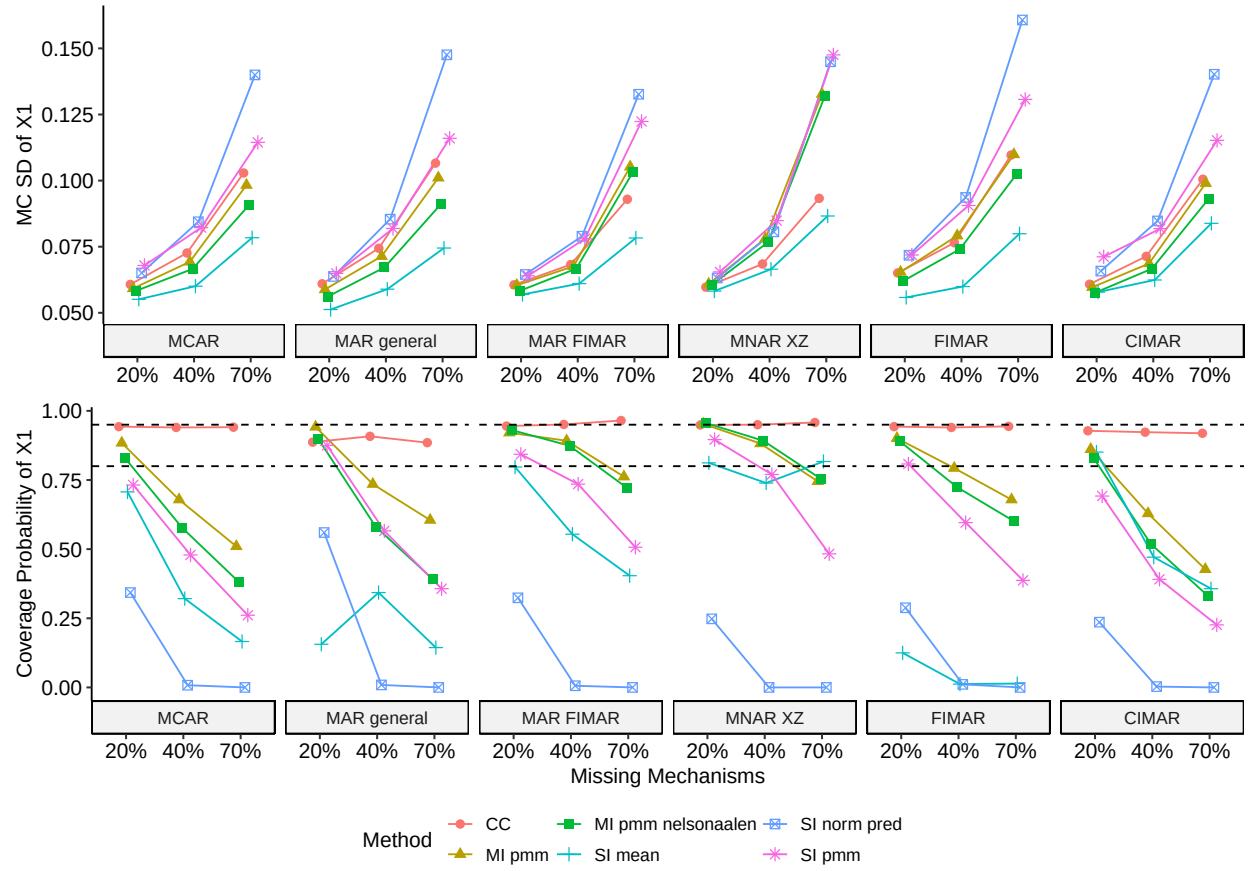


Figure 3.2: SD and Cov. prob. of X_1 in simulation 1. Dashed lines represent the 95% and 80% coverage probability.

Under traditional MAR, CC may lead to bias while MICE outperforms other methods both in terms of bias and efficiency. Under either MAR Z or MAR FIMAR, CC has 95% coverage probability and is unbiased since missing depend on either fully observed covariates or censoring time (Little and Rubin, 2019). However, MICE is more efficient compared with CC.

Under MNAR XZ, CC is valid since missingness depend only on covariates. MICE have relative large bias ($PB > 5$) while still efficient compared with CC. Under either MNAR FIMAR nor MNAR CIMAR, CC has smaller bias while is less efficient compared with MICE pmm or MICE pmm nelsenallen.

3.4.2 Simulation 2: multivariate missingness

In the second simulation setting, we considered multivariate missingness in order to mimic the realistic setting. $X_1, \dots, X_8, Z_1, Z_2$ were generated from a standard multivariate normal distribution with bivariate correlation $\rho = 0.5$. Survival times were simulated with hazard function $h(t|X, Z) = 0.09 \exp(X_1 + X_2 - X_3 - X_4 + 0.2X_5 + 0.2X_6 - 0.2X_7 - 0.2X_8 + Z_1 + Z_2)$. Censoring time were generated from exponential distribution with hazard 0.03 to give 35% censoring. Equal proportion of missingness was imposed on $X_1, \dots, X_8$ leading to 20%, 40% and 70% missingness and Z_1 and Z_2 were fully observed. Details of missing mechanisms can be found in Appendix.

Results are in Table 4. Under MCAR or when missingness depends on covariates only (observed or not), CC is unbiased. MICE would produce more efficient estimator. As the missing proportion increases, there is a general trend of increasing bias and decreasing precision for all methods above.

3.4.3 Summary

In summary, the results from the simulation study suggest the following:

1. CC analysis is unbiased when MCAR or when missingness depend on covariates, regardless of fully observed or not. Thus is recommended when missingness does not depend on outcomes.
2. MI produce more efficient estimators compared with CC while it is biased. MI is recommended when missing mechanism is general MAR.
3. As the missing proportion increases, all methods have increasing bias and decreasing precision.

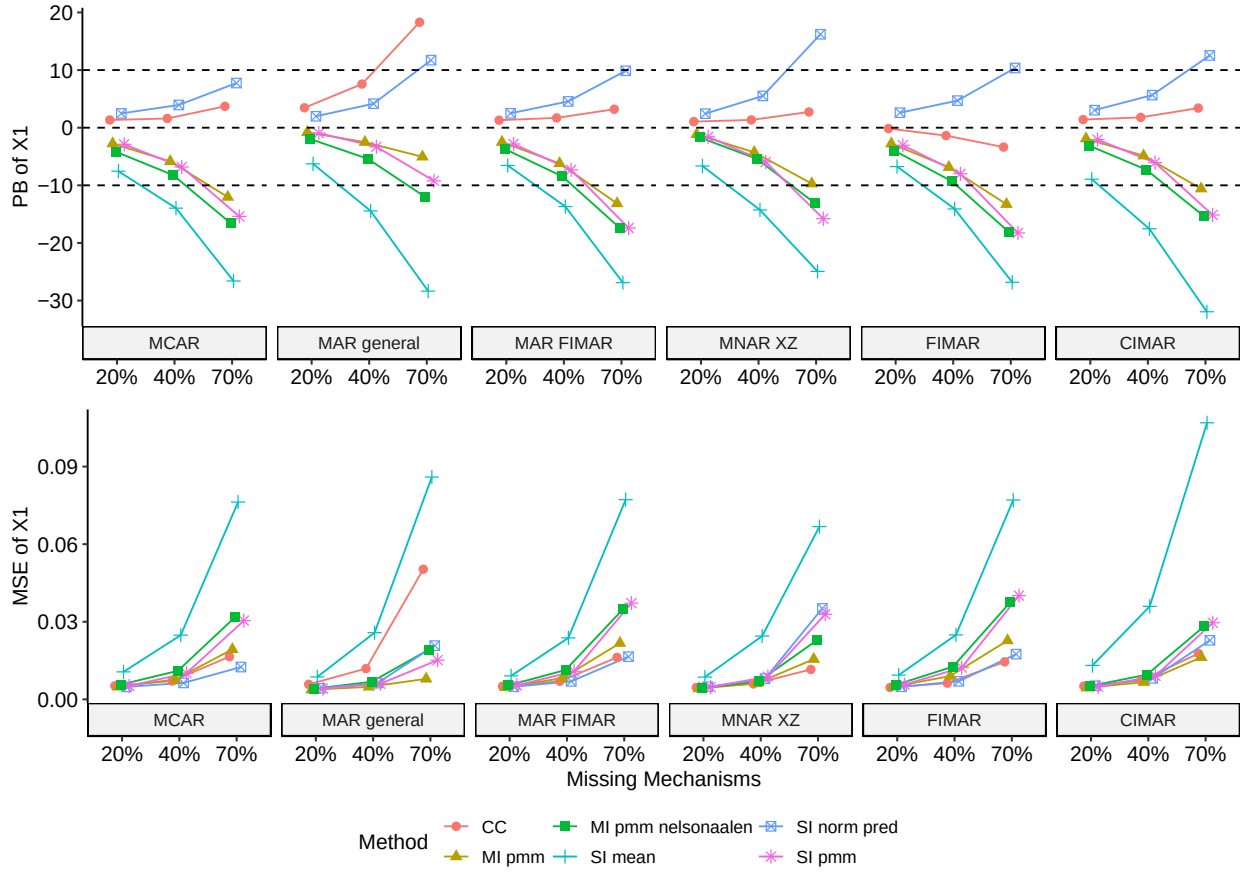


Figure 3.3: PB and MC MSE of X_1 in simulation 2. Dashed lines represent the $\pm 10\%$ PB.

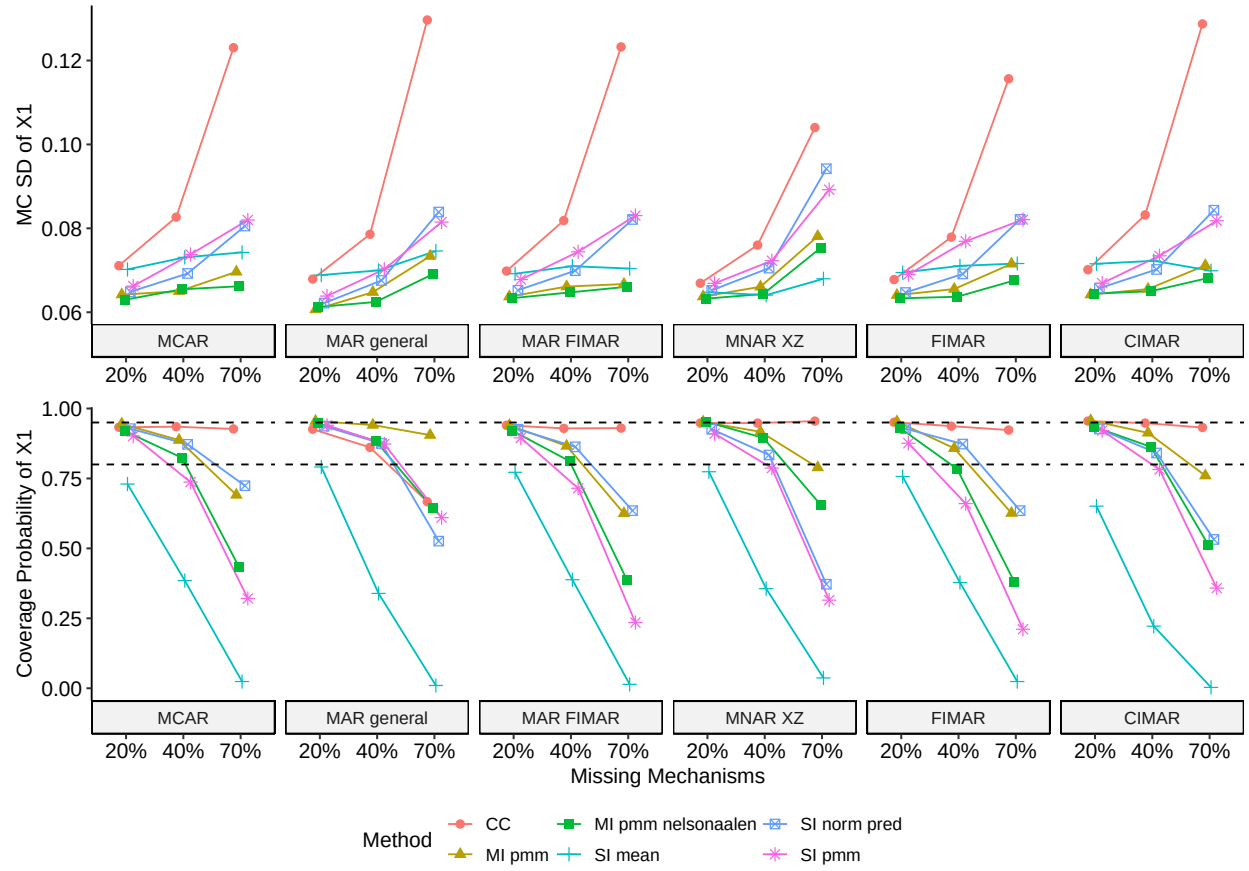


Figure 3.4: SD and Cov. prob. of X_1 in simulation 2. Dashed lines represent the 95% and 80% coverage probability.

3.5 Discussion

In this paper, we have discussed the identifiability of MAR and compared MICE with other commonly used methods when covariates are missing in Cox regression model. Based on the simulation results, three single imputation methods are not recommended because its large bias and "too small" standard error in most scenarios. While MICE is more efficient, CC analysis produce less biased and better coverage probability than MICE except when the missing mechanism is general MAR.

As pointed out in literature (Little, 1992), CC leads to valid inference when missingness depends only on covariates, regardless of observed or not. Therefore, CC is unbiased when missing mechanisms are MCAR and MNAR XZ. However, two MICE methods are mostly biased towards the null under different missing mechanisms and missing rate except when the data are MAR and has low missing rate. This is because the imputation is not entirely correct. MICE may impute covariates from models that are incompatible with the analysis model and lead to biased estimates. To avoid the model incompatibility, One can impute the missing covariates from a new model which contains information of analysis model and its density is proportion to the original imputation density (Bartlett et al., 2015). However, all imputation model are likely to be mis-specified to some extent, but biases may be small provided imputation models preserve those features of the data which are subsequently investigated. Therefore, auxiliary variables need to be included in order to solve model incompatibility and increase the plausibility of MAR.

In summarize, when data are general MAR, CC may lead to potential bias because missingness depends on the outcome. MICE is generally recommended because MAR assumption is satisfied. On the other hand, when data are not general MAR, MI should be cautiously used since both the MAR assumption is violated and model incompatibility problem exists. In practice, however, general MAR is untestable. It is possible that data are MNAR, FIMAR or CIMAR. This indicates that sensitivity analysis would be a possible way to evaluate the effects of missing covariates on MICE and complete case (CC) analysis may serve as a useful method for comparisons.

CHAPTER 4

Joint Models for Ordered Bivariate Survival times on Semi-Competing Risk Data

4.1 Introduction

Prostate cancer is one of the most common types of cancer in American men and early diagnosis is important to prolong cancer survival (American Cancer Society, 2020). Screening can increase early diagnosis rates and lead to early treatment. Randomized clinical trials, such as the PLCO (Prostate, Lung, Colorectal, and Ovarian) study, have been conducted to assess the effect of screening on cancer mortality. However, the early detection and prompt intervention enabled by screening do not necessarily result in significant reductions in cancer mortality due to issues such as overdiagnosis and overtreatment (Andriole et al., 2009, 2012). In this article, we are interested in testing the screening effect on cancer mortality, with the restriction that cancer diagnosis must occurs before cancer death.

Motivated by the idea that the time to the non-terminal event (cancer diagnosis) is strongly correlated with the time to the terminal event (death), A joint modelling approach is needed to capture the relationship between those two events and estimate the screening effect more efficiently.

A lot of methods have been developed based on the joint modeling idea, in survival analysis, this scenario is often referred to as semi-competing risk data (Fine et al., 2001), which involves a terminal event (e.g., death) and multiple non-terminal events (e.g., cancer diagnosis, recurrence). The terminal event censors the non-terminal event but not vice versa. Several joint modeling approaches for semi-competing risk data have been proposed to evaluate covariate effects, including copula-based methods (Chen, 2012; Emura and Chen, 2018) and illness-death models (Xu et al., 2010; Kalbfleisch and Prentice, 2011). However, these methods are not directly applicable to our setting because they do not assume an inherent ordering of events (e.g., death may occur before or after cancer recurrence). In our case, cancer diagnosis always precedes death. Gap time analysis has been employed to model

the relationship between the time to recurrence and the time from recurrence to death (Lin et al., 1999; Huang and Liu, 2007), automatically satisfying the order restriction. However, these methods focus on gap-specific covariate effects, whereas our objective is to assess the screening effect on marginal survival. Currently, there is limited literature addressing the order restrictions in semi-competing risk data (Wang, 2021).

In this article, we develop a novel joint model for ordered bivariate survival times of cancer diagnosis and cancer death to assess the screening effect on mortality using copulas. A copula is a bivariate distribution function that links two random variables via their marginal distributions by specifying the dependence structure (Emura and Chen, 2018). While extensive literature exists on copulas (see Nelsen (2006) for further details), traditional copulas do not guarantee the order restriction of random variables. In our case, cancer diagnosis must precede cancer death. There is a lack of research on copulas that explicitly consider order restrictions. We extend traditional copulas to scenarios with order restrictions and utilize the expanded copulas to construct the joint likelihood for ordered cancer diagnosis and cancer death for later inference. Simulation results demonstrate the nice finite-sample performance of the proposed method.

In Section 4.2, we introduce the data structure and basic notations. Section 4.3 extends traditional copulas to scenarios with order restrictions and presents the model setup. Section 4.4 describes the likelihood and estimation procedure. Section 4.5 evaluates the performance of the proposed method through simulation studies. A real data application is presented in Section 4.6, and a brief discussion concludes the paper in Section 4.7.

4.2 Notation and Data Structure

Let T_1^* denote the time to cancer diagnosis, and T_2^* denote the time to cancer mortality, respectively. Let $\mathbf{Z}$ denote a set of fully observed covariates. Let U denote the censoring time, which is independent of (T_1^*, T_2^*) given $\mathbf{Z}$. Additionally, the time to cancer diagnosis is assumed to occur before the time to cancer death, i.e., $T_1^* < T_2^*$ with probability 1. Let $(T_{1i}^*, T_{2i}^*, U_i, \mathbf{Z}_i)$, $i = 1, \dots, n$, be i.i.d. copies of $(T_1^*, T_2^*, U, \mathbf{Z})$. The observed data for subject i consist of $(T_{1i}, T_{2i}, \Delta_{1i}, \Delta_{2i}, \mathbf{Z}_i)$, where $T_{ki} = \min(T_{ki}^*, U_i)$, $k = 1, 2$, and $\Delta_{ki} = I(T_{ki} = T_{ki}^*)$, with $I(\cdot)$ denoting the indicator function.

4.3 Modified Copula and Model Specification

4.3.1 Copula With Order Restriction

Since we aim to utilize a copula to develop a joint model for cancer diagnosis and cancer death, we further assume that the joint distribution of T_1^* and T_2^* can be defined by a known copula with unknown parameters:

$$S(t_1, t_2) = \Pr(T_1^* > t_1, T_2^* > t_2) = C(S_{T_1^*}(t_1), S_{T_2^*}(t_2), \alpha), \quad t_1 \geq 0, t_2 \geq 0, \quad (4.1)$$

where $S(t_1, t_2)$ is the bivariate survival probability, $C(u, v, \alpha)$, $u, v \in [0, 1]$, is the specified copula, and $S_{T_k^*}$ is the marginal survival function for T_k^* , $k = 1, 2$.

However, the traditional bivariate copula, for example, the Clayton copula (Nelsen, 2006):

$$C(u, v, \alpha) = \Pr(U \leq u, V \leq v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha}, \quad \alpha > 0,$$

does not guarantee the order restriction of random variables. In our case, cancer diagnosis must precede cancer death. To extend the traditional copulas to scenarios with order restrictions, we first utilized the corollary of Sklar's theorem (Nelsen, 2006) in copula theory:

Theorem 4.1. *Let S be the joint survival function of two continuous random variables, and let F and G be the corresponding marginal survival functions. Let F^{-1} and G^{-1} be the quasi-inverses of F and G , respectively. For any $u, v \in [0, 1]$,*

$$C(u, v) = S(F^{-1}(u), G^{-1}(v)).$$

The theorem states that one can construct copulas from joint distribution functions and the corresponding marginal distributions.

To extend the copula to scenarios with order restrictions, we start with a joint distribution of two random variables, incorporating the order restrictions. The newly generated copula then contains the order information. By further substituting $u = S_{T_1^*}$ and $v = S_{T_2^*}$, similar to the approach used in traditional copulas, we can generate a new joint distribution where the new random variables T_1^* and T_2^* satisfy the specified order restrictions.

To illustrate, consider a toy example with order restrictions: we assume $Y \sim \text{Exponential}(\lambda)$ and $X | Y = y \sim \text{Uniform}(0, y)$. This toy example ensures $X < Y$ with probability 1. Therefore, applying Theorem 4.1 to this ordered bivariate joint distribution

yields the corresponding ordered copula (proof in Appendix C.1):

$$C(u, v) = \begin{cases} u, & \text{when } u < S_X \circ S_Y^{-1}(v), \\ (S_Y^{-1}(v) - S_X^{-1}(u)) f_X \circ S_Y^{-1}(v) + S_X \circ S_Y^{-1}(v), & \text{when } u \geq S_X \circ S_Y^{-1}(v), \end{cases} \quad (4.2)$$

where $S_X(x) = \exp(-\lambda x) - \lambda x E_1(\lambda x)$ and $S_Y(y) = \exp(-\lambda y)$ are the marginal survival functions of random variables X and Y , respectively, in the toy example, $E_1(x) = \int_x^\infty \frac{e^{-t}}{t} dt$, and $f_X(x) = \lambda E_1(\lambda x)$ is the PDF of X .

We are now ready to use the derived “ordered” copulas to generate joint distributions of the target random variables T_1^* and T_2^* based on (4.1). Equation (4.1) states that one can generate the joint distribution by plugging two arbitrary marginal survival functions into a known copula. It is natural to ask what assumption is needed to ensure that the new joint survival function of T_1^* and T_2^* is not only a legitimate joint distribution but also a joint distribution that satisfies the order restriction, i.e., $T_1^* < T_2^*$ with probability 1. Here is the restriction (proof in Appendix C.2) that must be satisfied:

$$P(t) = Q(t), \quad \forall t \in [0, \infty), \quad (4.3)$$

where $P(t) = S_Y^{-1} \circ S_{T_2^*}(t)$ and $Q(t) = S_X^{-1} \circ S_{T_1^*}(t)$.

4.3.2 Model Specification

Suppose that the marginal distribution of T_2^* follows a Cox proportional hazards model:

$$\Lambda_{T_2^*}(t \mid \mathbf{z}) = \eta_2(\mathbf{z}) H_2(t), \quad (4.4)$$

where $\Lambda_{T_2^*}(t \mid \mathbf{z})$ is the cumulative hazard of T_2^* given covariates $\mathbf{z}$, $\eta_2(\mathbf{z}) = \exp(\beta \mathbf{z})$, and $H_2(t)$ is the baseline cumulative hazard for time to cancer mortality.

By the restriction (4.3), the marginal distribution of T_1^* is given by

$$S_1(t) = S_X \circ P(t) = S_X \left(\frac{H_2(t) \eta_2}{\lambda} \right).$$

Then, copula (4.2) yields the joint distribution function of T_1^* and T_2^* with order restrictions

as:

$$\begin{aligned}
S(t_1, t_2) &= \Pr(T_1^* > t_1, T_2^* > t_2) \\
&= \begin{cases} \Pr(T_1^* > t_1) = S_{T_1^*}(t_1), & \text{when } t_1 > M(t_2), \\ (P(t_2) - P(t_1))f_X \circ P(t_1) + S_X \circ P(t_2), & \text{when } t_1 \leq M(t_2), \end{cases} \quad (4.5)
\end{aligned}$$

where $M(t_2) = S_{T_1^*}^{-1} \circ S_X \circ S_Y^{-1} \circ S_{T_2^*}(t_2)$.

4.4 Likelihood and Estimation

4.4.1 Likelihood Construction

We now utilize (4.5) to construct the joint likelihood of cancer diagnosis and time to death with semi-competing risk data $(T_{1i}, T_{2i}, \Delta_{1i}, \Delta_{2i}, Z_i), i = 1, \dots, n$. The likelihood for a single subject belongs to the following scenarios:

1. Subject has diagnosis at T_1 and death at T_2 ($\Delta_1 = \Delta_2 = 1$):

$$\begin{aligned}
L_{11} &= \Pr(T_1^* = t_1, T_2^* = t_2) \\
&= \begin{cases} 0, & \text{when } t_1 > M(t_2), \\ \frac{\partial P(t_1)}{\partial t_1} (-f'_X(P(t_2))) \frac{\partial P(t_2)}{\partial t_2}, & \text{when } t_1 \leq M(t_2), \end{cases}
\end{aligned}$$

where

$$\frac{\partial P(t_1)}{\partial t_1} = \frac{\eta_2 dH_2(t_1)}{\lambda dt_1}, \quad (-f'_X(x)) = \lambda \exp(-\lambda x) \frac{1}{x}.$$

Since $T_1^* < T_2^*$ with probability 1, it follows that $t_1 < t_2$, which implies $Q(t_1) = P(t_1) < P(t_2)$. This leads to $t_1 \leq M(t_2)$. Thus, the likelihood simplifies to:

$$L_{11} = \frac{\partial P(t_1)}{\partial t_1} (-f'_X(P(t_2))) \frac{\partial P(t_2)}{\partial t_2}.$$

2. Subject has diagnosis at T_1 and is censored at T_2 ($\Delta_1 = 1, \Delta_2 = 0$):

$$L_{10} = \Pr(T_1^* = t_1, T_2^* > t_2) = \frac{\partial P(t_1)}{\partial t_1} f_X \circ P(t_2).$$

3. Subject is censored before any event occurs ($\Delta_1 = 0, \Delta_2 = 0, T_1 = T_2$):

$$L_{00} = \Pr(T_1^* > t_1, T_2^* > t_2) = S_{T_1^*}(t_1).$$

Therefore, the joint log-likelihood for a single subject is:

$$l_i = \Delta_{1i}\Delta_{2i} \log L_{11}(t_{1i}, t_{2i}) + \Delta_{1i}(1 - \Delta_{2i}) \log L_{10}(t_{1i}, t_{2i}) + (1 - \Delta_{1i}) \log L_{00}(t_{1i}, t_{2i}).$$

Alternatively, we can express the log-likelihood as $l_i = l_{1i} + l_{2i}$, where

$$l_{1i} = \Delta_{1i} \log f_{T_1^*}(t_{1i}) + (1 - \Delta_{1i}) \log S_{T_1^*}(t_{1i}),$$

$$l_{2i} = \Delta_{1i}\Delta_{2i} \log [f_{T_2^*|T_1^*}(t_{2i} | t_{1i})] + \Delta_{1i}(1 - \Delta_{2i}) \log [S_{T_2^*|T_1^*}(t_{2i} | t_{1i})].$$

The i th subject's likelihood is divided into two parts, where l_{1i} is based on the information from time zero to cancer diagnosis, while l_{2i} is based on the information from cancer diagnosis to cancer mortality.

4.4.2 Estimation

Our parameters of interest are $\Omega = (\lambda, \beta, H_2)$, where λ is the parameter introduced in the toy example, β represents the effects of covariates on T_2^* , and the baseline cumulative hazard function H_2 is modeled as a step function with jumps dH_2 occurring only at the observed cancer diagnosis and cancer death times.

Since the parameters include both finite-dimensional components (λ, β) and an infinite-dimensional component (H_2) , we employ the EM algorithm (Tsodikov, 2003; Rice and Tsodikov, 2017) and profile likelihood (Murphy and Van der Vaart, 2000) to obtain the nonparametric maximum likelihood estimator (NPMLE) for Ω .

4.4.2.1 Martingale Theory

Let $N_{ki}(t) = I(T_{ki} \leq t, \Delta_{ki} = 1)$, $k = 1, 2$, denote subject i 's observed counting process for cancer diagnosis and death, respectively. Let $Y_{ki}(t) = I(T_{ki} \leq t)$, $k = 1, 2$, denote subject i 's at-risk process for cancer diagnosis and death, respectively.

The log-likelihood for subject i can be expressed in terms of counting processes as:

$$\begin{aligned} l_i &= l_{1i} + l_{2i} \\ &= \int_0^\infty [\log d\Lambda_{1i}(t) dN_{1i}(t) - Y_{1i}(t) d\Lambda_{1i}(t)] \\ &\quad + \int_0^\infty \int_{t_1}^\infty [\log d\Lambda_{2i}(t | t_1) dN_{2i}(t) - Y_{2i}(t) d\Lambda_{2i}(t)] dN_{1i}(t_1). \end{aligned}$$

The martingales $dM_{1i}(t)$ and $dM_{2i}(t | t_1)$ are derived from the observed counting processes

relative to the filtration $\mathcal{F}_i(t_-) = \sigma\{N_{1i}(s), N_{2i}(s), Y_{1i}(s), Y_{2i}(s), z_i : s \in [0, t]\}$, and are defined as:

$$\begin{aligned} dM_{1i}(t) &= dN_{1i}(t) - Y_{1i}(t) d\Lambda_{1i}(t), \\ dM_{2i}(t | t_1) &= dN_{2i}(t) dN_{1i}(t_1) - Y_{2i}(t) dN_{1i}(t_1) d\Lambda_{2i}(t | t_1). \end{aligned}$$

where

$$\begin{aligned} d\Lambda_{1i}(t) &= \Pr(dN_{1i}(t) = 1 | Y_{1i}(t) = 1) \\ &= \frac{\Pr(T_{1i}^* = t, T_{2i}^* > t)}{\Pr(T_{1i}^* \geq t, T_{2i}^* > t)} \\ &= \frac{\frac{\partial P_i(t)}{\partial t} f_X \circ P_i(t)}{S_X \circ P_i(t)} \\ &\triangleq \Theta_{1i}(t), \\ d\Lambda_{2i}(t | t_1) &= \Pr(dN_{2i}(t | t_1) = 1 | Y_{2i}(t | t_1) = 1) \\ &= \frac{\Pr(T_{1i}^* = t_1, T_{2i}^* = t)}{\Pr(T_{1i}^* = t_1, T_{2i}^* \geq t)} \\ &= \frac{[P_i(t) - P_i(t_1)] f_X \circ P_i(t) + S_X \circ P_i(t)}{\frac{\partial P_i(t_1)}{\partial t_1} f_X \circ P_i(t)} \\ &\triangleq \Theta_{2i}(t | t_1). \end{aligned}$$

Here, $d\Lambda_{1i}(t)$ denotes the hazard of subject i being diagnosed at time t , while $d\Lambda_{2i}(t | t_1)$ denotes the hazard of subject i experiencing cancer mortality at time t , conditional on being diagnosed at time t_1 .

4.4.2.2 Score Function

Due to the non-identifiability problem of λ , we fix $\lambda = 1$ for simplicity. In fact, the data do not contain any information about λ (proof provided in Appendix C.3). Therefore, we define the partial derivatives of $\Theta_{1i}(t)$ and $\Theta_{2i}(t; t_1)$ with respect to the remaining parameters of interest, $\{dH_2(s)\}$ and β , as follows:

$$\begin{aligned} \dot{\Theta}_{1i, dH_2(s)}(t) &= \frac{\partial \Theta_{1i}(t)}{\partial dH_2(s)}, \\ \dot{\Theta}_{1i, \beta}(t) &= \frac{\partial \Theta_{1i}(t)}{\partial \beta}, \\ \dot{\Theta}_{2i, dH_2(s)}(t; t_1) &= \frac{\partial \Theta_{2i}(t; t_1)}{\partial dH_2(s)}, \\ \dot{\Theta}_{2i, \beta}(t; t_1) &= \frac{\partial \Theta_{2i}(t; t_1)}{\partial \beta}. \end{aligned}$$

As in Hu and Tsodikov (2014), for a functional $J(f)$, $f = f(x)$, the functional derivative in the above equations is defined as

$$\frac{\partial J(f)}{\partial f(s)} = \left. \frac{\partial J(f + \epsilon g)}{\partial \epsilon} \right|_{\epsilon=0, g=1\{r>s\}}.$$

Applying the functional derivative to the full log-likelihood, with respect to the infinite-dimensional parameters $\{dH_2(s)\}$, we can obtain the score function for $\{dH_2(s)\}$ as (proof in Appendix C.1).

$$U_{dH(s)} = \sum_i \left\{ \int_s^\tau \frac{\dot{\Theta}_{1i,dH(s)}(t)}{\Theta_{1i}(t)} dM_{1i}(t) + \int_0^\tau \int_s^t \frac{\dot{\Theta}_{2i,dH(s)}(t; t_1)}{\Theta_{2i}(t; t_1)} dM_{2i}(t; t_1) + \int_0^s \frac{dM_{2i}(s; t_1)}{dH(s)} \right\},$$

Taking derivative of the log-likelihood, with respect to the regression parameter β , we can have the score function for β as

$$U_\beta = \sum_i \left\{ \int_0^\tau \frac{\dot{\Theta}_{1i,\beta}(t)}{\Theta_{1i}(t)} dM_{1i}(t) + \int_0^\tau \int_0^t \frac{\dot{\Theta}_{2i,\beta}(t; t_1)}{\Theta_{2i}(t; t_1)} dM_{2i}(t; t_1) \right\},$$

which is also a martingale under the true model.

Set score functions (2.7) and (2.8) to be zero, and solve them, can give the NPMLE $\hat{\Omega} = (\hat{\beta}, \{d\hat{H}\})$.

4.4.2.3 Estimation

The NPMLE can be obtained by directly maximizing the log-likelihood derived in Section 4.4.2.1. However, straightforward Newton-type methods are computationally expensive, primarily due to the nonparametric components $\{dH_2\}$, which are of infinite dimension. Instead, we apply the profile likelihood and EM algorithm to jointly estimate β and $\{dH_2\}$.

The algorithm follows a nested procedure. Given fixed β , we first obtain the estimators of $\{dH_2\}$ using the EM algorithm. We then calculate the profile likelihood $l_{pr} = l(\beta, \{d\hat{H}_2(\beta)\})$ by substituting the updated $\{dH_2\}$. Finally, β is estimated by maximizing the profile likelihood.

The key step is to estimate $\{dH_2\}$, given β . We employ the EM algorithm to estimate the cumulative baseline hazards (proof provided in Appendix C.1). The score functions for $\{dH_2\}$ in the $(k + 1)$ th iteration are given by:

$$U_{dH_2^{(j+1)}}(s) = \sum_{i=1}^n \left\{ \frac{dN_{1i}(s) + dN_{2i}(s)}{dH_2^{(j+1)}(s)} - \Psi_i^{(j)}(s) + \left[\frac{dH_2^{(j)}(s)}{dH_2^{(j+1)}(s)} - 1 \right] \theta_i^{(j)}(s) \right\} = 0, \quad (4.6)$$

where

$$\begin{aligned} \Psi^{(j)}(s) = & \Delta_1 \Delta_2 \left(1 + \frac{1}{\lambda P(t_1)} \right) \eta_2 Y_2(s) \\ & + \Delta_1 (1 - \Delta_2) \left\{ Y_2(s) G_1(t_2, \infty) - [1 - Y_2(s)] [G_{2,t_2s} - G_1(s, \infty)] \right\} \\ & + (1 - \Delta_1) (1 - \Delta_2) \left\{ Y_2(s) J_1(t_2, \infty) - [1 - Y_2(s)] [J_{2,ts} + J_{3,ts} - J_1(s, \infty)] \right\}, \end{aligned}$$

$$\theta^{(j)}(s) = \Delta_1 (1 - \Delta_2) [1 - Y_2(s)] G_{2,t_2s} + (1 - \Delta_1) (1 - \Delta_2) [1 - Y_2(s)] (J_{2,ts} + J_{3,ts}).$$

Here,

$$\begin{aligned} G_1(a, b) &= \int_a^b G_{1,t_2m} dH_2(m), \\ G_{1,t_2m} &= \frac{\eta_2^2 \left[1 + \frac{1}{\lambda P(t_1)} \right] [-f'_X(P(m))]}{\lambda f_X \circ P(t_2)}, \\ G_{2,t_2s} &= \frac{[-f'_X(P(s))] \eta_2}{\lambda f_X \circ P(t_2)}, \\ J_1(a, b) &= \int_a^b J_{1,tm} dH_2(m), \\ J_{1,tm} &= \frac{\left(1 + \frac{1}{\lambda P(t_1)} \right) \eta_2^2 [P(m) - P(t_1)] [-f'_X(P(m))]}{S_X \circ P(t_1) \lambda}, \\ J_{2,ts} &= \frac{f_X \circ P(s) \eta_2}{S_X \circ P(t_1) \lambda}, \\ J_{3,ts} &= \frac{[P(s) - P(t_1)] [-f'_X(P(s))] \eta_2}{S_X \circ P(t_1) \lambda}. \end{aligned}$$

Solving the equation, we obtain Breslow-type estimators:

$$dH_2^{(j+1)}(s) = \frac{\sum_i (dN_{1i}(s) + dN_{2i}(s)) + \left[\sum_i \theta_i^{(j)}(s) \right] dH_2^{(j)}(s)}{\sum_i \left[\Psi_i^{(j)}(s) + \theta_i^{(j)}(s) \right]}.$$

$dH_2(s)$ are updated iteratively, $k = 0, 1, 2, \dots$, until convergence.

The estimation procedure is outlined as follows:

Starting with an initial value $\beta^{(0)}$, set the iteration counter to $j = 0$.

1. Given $\beta = \beta^{(j)}$, estimate $\{dH_2\}$ using the EM algorithm:
 - (a) Set the EM iteration counter to $k = 0$ and initialize $d\hat{H}_2^{(0)}(s)$ such that all jumps of cancer diagnosis and cancer death in the baseline hazards have equal size
 - (b) With β fixed, update $d\hat{H}_2^{(k+1)}(s)$ using equation (4.6).
 - (c) Repeat step (b) until convergence, i.e.,

$$\left\| d\hat{H}_2^{(k+1)}(s) - d\hat{H}_2^{(k)}(s) \right\|_2 < \epsilon.$$

2. Given the converged $\{dH_2\}$, maximize the profile log-likelihood $l_{pr}((\beta))$ with respect to (β) using a numerical optimization method, such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm implemented in the ‘optim’ function in R.

Repeat steps 1-2 until convergence of $l_{pr}((\beta))$:

$$|l_{pr}((\beta)^{(j+1)}) - l_{pr}((\beta)^{(j)})| < \xi.$$

Note the convergence tolerance for the inner loop (EM algorithm used to estimate baseline hazards, given (β)) should be stricter than that for the outer loop, e.g., $\epsilon = 10^{-6}, \xi = 10^{-5}$.

4.5 Simulation Studies

We conducted simulation studies using 1000 Monte Carlo datasets to evaluate the performance of the proposed method. A simple scenario with two covariates, Z_1 and Z_2 , was considered, where $Z_1 \sim \text{Normal}(0, 0.16)$ and $Z_2 \sim \text{Bernoulli}(0.5)$. Marginal survival times for cancer mortality were simulated using the hazard function $h_2(t) = 2t \exp(\beta_1 z_1 + \beta_2 z_2)$, with true parameters $(\beta_1, \beta_2) = (-0.5, 0.3)$. Marginal survival times for cancer diagnosis were simulated using $S_{T_1^*}(t) = S_X \left[\frac{\eta_2 H_2(t)}{\lambda} \right]$. For simplicity, λ was fixed at 1, as in the toy example. Censoring times U were simulated from an exponential distribution with a hazard rate of 0.65, resulting in approximately 25% of intermediate events and 40% of terminal events being censored.

4.5.1 Finite-sample Properties of Parameter Estimates

We first evaluated the finite-sample performance of the proposed model using sample sizes of 100, 200, and 400. Standard errors were computed via numerical approximation of the Hessian matrix at the solution.

Table 4.1 summarizes the results. Both β_1 and β_2 demonstrate negligible bias across all sample sizes (100, 200, and 400). As expected, the bias and standard error decrease with increasing sample size. The coverage probabilities of the corresponding confidence intervals are close to the nominal 95% level. Furthermore, larger sample sizes lead to better agreement between the asymptotic standard errors and empirical standard deviations, confirming the reliability of the proposed estimators.

Table 4.1: Simulation results for the proposed model

N	β	Truth	MC Bias	AVE.SE	MC SD	Cov. prob
100	β_1	-0.5	0.029	0.845	0.805	0.943
	β_2	0.3	0.002	0.256	0.254	0.941
200	β_1	-0.5	-0.014	0.581	0.556	0.938
	β_2	0.3	-0.002	0.187	0.177	0.941
400	β_1	-0.5	-0.021	0.397	0.392	0.949
	β_2	0.3	-0.004	0.127	0.125	0.944

AVE.SE: average of estimated standard errors

MC SD: Monte Carlo standard deviation

Cov. prob.: the coverage probability of a 95% Wald confidence interval

4.5.2 Hypothesis Testing

We conducted simulations to construct the likelihood ratio test (LRT) for assessing the effect of the binary covariate on cancer mortality. The simulation setup was the same as described earlier, and the sample size was fixed at 400. Under the null hypothesis, we set $\beta_2 = 0$. The LRT test statistic was computed based on the proposed model, with and without the inclusion of the binary covariate. Empirical p -values were obtained through 1000 Monte Carlo replicates. The results indicate that the 0.05 significance level is well maintained (0.049).

4.6 Data Application

We applied the proposed method to the motivating data from the PLCO trial, focusing on testing the effect of screening on cancer mortality. This trial randomized participants (age 55 – 74 at randomization) into two arms: the screening arm and the control arm. The overall sample size was 76,674 subjects, with 38,335 in the screening arm and 38,339 in the control arm. In the screening arm, 4,418 subjects (11.52%) were diagnosed with prostate cancer, and 145 (0.38%) died of it, while in the control arm, 4,036 subjects (10.53%) were diagnosed with prostate cancer, and 142 (0.37%) died of it. The maximum follow-up time

was 13 years. In addition to the screening effect, we were also interested in studying the effect of age, which ranged from 55 to 74 years.

The parameter estimates of interest are presented in Table 4.2 for both the proposed joint model and the Cox model for marginal death. Based on the estimates from our proposed joint model, older patients had a greater risk of death ($\beta = 0.116$, HR = 1.12, p-value < 0.0001), while screening did not significantly reduce cancer mortality ($\beta = 0.034$, HR = 1.03, p-value = 0.758).

The point estimates were similar between the proposed model and the Cox model, as expected, given the large sample size and the unbiased nature of both methods. However, the standard errors for the β coefficients in the proposed method were slightly smaller than those from the marginal models. This is because the proposed joint model simultaneously incorporates information from both cancer diagnosis and cancer death, thereby improving efficiency. Unlike the simple Cox model for marginal death, our proposed model provides a more comprehensive view of the disease progression process by incorporating incidence data.

Table 4.2: Estimations from the proposed model and Cox model for cancer death

Parameter	Proposed Joint Model			Cox for Marginal Death		
	Estimates	SE	p-value	Estimates	SE	p-value
Age	0.116	0.011	0.000	0.117	0.011	0.000
Screen vs. not	0.034	0.112	0.758	0.035	0.118	0.766

4.7 Discussion

This study introduces a novel copula-based joint modeling framework for analyzing the relationship between cancer diagnosis and cancer mortality, with a primary focus on testing the effect of screening on cancer mortality. The proposed method extends traditional copulas to accommodate order restrictions, ensuring the natural sequence of cancer diagnosis preceding cancer death. This expanded copula framework allows for explicit modeling of the dependence between diagnosis and death while incorporating covariate effects and leveraging additional information from intermediate events (cancer diagnosis) to improve estimation efficiency.

The proposed framework offers several advantages. First, it provides unbiased and efficient estimates by additionally incorporating diagnostic information into the modeling of terminal events (death). Second, the generation of a copula with order restrictions introduces a new approach for analyzing bivariate survival data, offering significant potential for applications

in clinical trials, epidemiological studies, and other fields where intermediate and terminal events are of interest.

Despite its strengths, the method also has limitations. The current implementation assumes a fixed copula structure, which may not fully capture all potential dependence patterns between sequential random variables. Future research could explore generalized toy examples to extend the methodology.

CHAPTER 5

Conclusion

This dissertation develops and evaluates methods for some common statistical challenges arising from clinical and epidemiological studies.

In Chapter II, we develop statistical methods for robust covariate adjustment to improve efficiency of inferences in randomized clinical trials when covariates are subject to missingness. We propose three different implementation strategies when multiple baseline covariates are likely missing. These strategies can be viewed as augmented methods with different ways to build the working models and handle the high dimensional sets of ‘covariates’. Notably, we find that knowledge from missing data literature on imputation does not necessarily apply in this setting. For instance, missing at random (MAR) assumption is not required, and the imputation model should not include the outcome. Multiple imputation has good performance but is not better than regression imputation.

Without assuming the data are MAR, Our findings indicate that the proposed methods yield consistent and asymptotically normal estimators of the treatment effects, provided the treatment assignment is independent of the missingness. Furthermore, the proposed estimators are guaranteed to be more efficient than the unadjusted estimator and are robust to model misspecifications.

In Chapter III, we discuss the plausibility of the MAR assumption in the context of survival analysis. We conducted extensive simulation studies to evaluate the performance of several popular methods, including multiple imputation by chained equations (MICE), complete-case analysis (CC), and single imputation (SI), under various missingness mechanisms and levels of missingness, with a particular focus on two alternative mechanisms to MAR.

The simulation results suggest that three single imputation methods are not recommended. CC leads to valid inference when missingness depends only on covariates, regardless of observed or not. Although commonly used in practice, the performance of MICE is unsatisfactory in terms of both bias and variance. In many settings, MICE does not exhibit a

clear advantage over CC. Based on these results, we recommend that CC analysis be included as a sensitivity analysis alongside MICE in practical applications.

In Chapter IV, we extend traditional copulas to scenarios with order restrictions and utilize the derived copulas to develop a novel joint model for ordered bivariate survival times of cancer diagnosis and cancer death to assess the screening effect on mortality. This expanded copula framework explicitly models the dependence between diagnosis and death, leveraging additional information from intermediate events (cancer diagnosis) to improve estimation efficiency.

Simulation results demonstrate the strong finite-sample performance of the proposed method. Additionally, the proposed framework introduces an innovative approach for analyzing bivariate survival data, offering significant potential for applications in clinical trials, epidemiological studies, and other fields where intermediate and terminal events are of interest.

APPENDIX A

Appendix Chapter 2

A.1 Equivalence between the estimand and expectation difference of the outcome

Stable unit treatment value assumption (SUTVA): The observed outcome is equivalent to the potential outcome of that treatment, regardless of how the treatment assignment is administered or the potential outcomes of other individuals. This can be expressed as $Y_i = Y_i^1 A_i + Y_i^0 (1 - A_i)$. In randomized clinical trials, we have:

$$\begin{aligned} E(Y_i^1) &= E(Y_i^1 | A_i = 1) \text{ by } A \perp\!\!\!\perp (X, Z, Y^0, Y^1) \\ &= E(A_i Y_i^1 + (1 - A_i) Y_i^0 | A_i = 1) \\ &= E(Y_i | A_i = 1) \text{ by SUTVA.} \end{aligned}$$

Follow the same technique, we can show that $E(Y_i^0) = E(Y_i | A_i = 0)$. Therefore, we have $\theta = E(Y_i^1) - E(Y_i^0) = E(Y_i | A_i = 1) - E(Y_i | A_i = 0)$.

A.2 Consistency of estimand using different imputation methods

Suppose we have only two covariates X_1 and X_2 and R is the observation indicator for X_1 . Concerning the $\hat{\theta}_{augimp}$ in (2.8), $\frac{\sum_{i=1}^n A_i Y_i}{\sum_{i=1}^n A_i} - \frac{\sum_{i=1}^n (1 - A_i) Y_i}{\sum_{i=1}^n (1 - A_i)} = \bar{Y}_1 - \bar{Y}_0$ is consistent for θ . Because $\hat{\pi} = n_1/n \xrightarrow{p} \pi$ and $q(\hat{\zeta}_a, g(X_{1i}^{imp}, X_2)) \xrightarrow{p} q_a(g(X_{1i}^{imp}, X_2)), a = 0, 1$, the augmentation term is approximately equal to $\frac{1}{n} \sum_{i=1}^n \{(A_i - \pi)[(1 - \pi)^{-1} q_0(g(X_{1i}^{imp}, X_2)) + \pi^{-1} q_1(g(X_{1i}^{imp}, X_2))]\} \xrightarrow{p}$

$E[(A - \pi)[(1 - \pi)^{-1}q_0(g(X_{1i}^{imp}, X_2)) + \pi^{-1}q_1(g(X_{1i}^{imp}, X_2))]$. When we use the mean value of X_1 or X_2 to impute X_1 , $q_a(g(X_{1i}^{imp}, X_2)), a = 0, 1$ is a function of X_1, X_2, R . Under the assumption: $A \perp\!\!\!\perp (X_1, X_2, R)$, the augmentation term converges to 0. When we include the outcome in the imputation model, $q_a(g(X_{1i}^{imp}, X_2)), a = 0, 1$ now becomes a function of X_1, X_2, Y, R . However, Y is not independent of A , which introduces the potential for a non-zero value in the augmented term in (2.8). Therefore, imputation using the outcome may lead to an inconsistent estimator.

A.3 Imputation as a two-step augmented method

For simplicity, we will continue to consider the setting in Appendix A.2. We will illustrate the statement for $A = 1$, as $A = 0$ follows the same technique.

Suppose we first conduct mean imputation for X_1 and then apply the augmented method to the imputed data set. We build the imputation model $X_1^{mimp} = RX_1 + (1 - R)\frac{\sum_{i=1}^n R_i X_{1i}}{\sum_{i=1}^n R_i}$ and working model for $E(Y|X_1^{mimp}, X_2, A = 1)$ as $q_{mimp}(X_1^{mimp}, X_2, \gamma) = \gamma_0 + \gamma_1 X_1^{mimp} + \gamma_2 X_2$, where $\gamma = (\gamma_0, \gamma_1, \gamma_2)^T$. Thus the estimator for outcome in the treatment group is given by: $\hat{\mu}_1^{mimp} = \frac{1}{n} \sum_{i=1}^n \{ \frac{A_i Y_i}{\hat{\pi}} - (A_i - \hat{\pi})\hat{\pi}^{-1} \{ \hat{\gamma}_0 + \hat{\gamma}_1 [R_i X_{1i} + (1 - R_i)\frac{\sum_{i=1}^n R_i X_{1i}}{\sum_{i=1}^n R_i}] + \hat{\gamma}_2 X_{2i} \} \}$. This is asymptotically equivalent to:

$$\frac{1}{n} \sum_{i=1}^n \{ \frac{A_i Y_i}{\hat{\pi}} - (A_i - \hat{\pi})\hat{\pi}^{-1} \{ (\gamma_0^* + \gamma_1^* E(X_1|R = 1)) + \gamma_1^* R_i X_{1i} - \gamma_1^* E(X_1|R = 1)R_i + \gamma_2^* X_{2i} \} \} \quad (\text{A.1})$$

by using $\gamma \xrightarrow{p} \gamma^*$ and $\frac{\sum_{i=1}^n R_i X_{1i}}{\sum_{i=1}^n R_i} \xrightarrow{p} E(X_1|R = 1)$. Note that γ^* is a constant vector, thus $q(\mathbf{g}(\mathbf{X}_i^{\text{imp}})) = (\gamma_0^* + \gamma_1^* E(X_1|R = 1)) + \gamma_1^* R_i X_{1i} - \gamma_1^* E(X_1|R = 1)R_i + \gamma_2^* X_{2i}$ actually is a function of $X_1 R, R, X_2$ and falls into the framework in Section 2.4.3. It's worth noting that regression imputation with covariates only, followed by applying the augmented method to the imputed dataset, can also be viewed as a two-step augmented method.

A.4 Comparison of imputation method and modified imputation method

We will illustrate this point using the first simulation setting except that now the R and ϵ are generated by $R \sim \text{Bernoulli}(p)$ and $\epsilon|R = r \sim N(m_r, 1), r = 0, 1$. We will illustrate the calculation for $A = 1$, as $A = 0$ follows the same technique.

For X imputation first, then the augmented method. Let $X_1^{ximp} = RX_1 + (1 - R)E(X_1|X_2, R = 1)$, we could build the imputation model for $E(X_1|X_2, R = 1)$ as $p(X_2, \eta) = \eta_0 + \eta_1 X_2$, where $\eta = (\eta_0, \eta_1)^T$ and working model for $E(Y|X_1^{ximp}, X_2, A = 1) = q_{ximp}(X_1^{ximp}, X_2, \delta) = \delta_0 + \delta_1 X_1^{ximp} + \delta_2 X_2$, where $\delta = (\delta_0, \delta_1, \delta_2)^T$. Let $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 A + \beta_4 X_2 A + \epsilon$. To obtain the optimal augmented estimator, we need to calculate the true form of $E(Y|A = a, X_1 R, X_2, R), a = 0, 1$. our optimal $E(Y|A = 1, X_1 R, X_2, R) = RE(Y|A = 1, X_1, X_2, R = 1) + (1 - R)E(Y|A = 1, X_2, R = 0) = (\beta_0 + \beta_3 + m_0 + \beta_1 \eta_0^*) + \beta_1 X_1 R + (\beta_2 + \beta_4 + \beta_1 \eta_1^*) X_2 + (m_1 - m_0 - \beta_1 \eta_0^*) R - \beta_1 \eta_1^* X_2 R$, where η^* is the converged value of $\hat{\eta}$. Thus, $\hat{\mu}_1^{ximp}$ is asymptotically equivalent to

$$\frac{1}{n} \sum_{i=1}^n \frac{A_i Y_i}{\hat{\pi}} - (A_i - \hat{\pi}) \hat{\pi}^{-1} \{(\delta_0^* + \delta_1^* \eta_0^*) + \delta_1^* R_i X_{1i} - \delta_1^* \eta_0^* R_i + (\eta_1^* \delta_1^* + \delta_2^*) X_{2i} - \eta_1^* \delta_1^* R_i X_{2i}\}, \quad (\text{A.2})$$

where $\delta \rightarrow \delta^*$. We could see that there is a restriction on the coefficient of R in (A.2), leading to the potential loss of efficiency for θ . However, if we apply the modified imputation method. Suppose we build the working model for X imputation + augment (with R) as $q_{ximpr}(X_1^{ximpr}, X_2, R, \kappa) = \kappa_0 + \kappa_1 X_1^{ximpr} + \kappa_2 X_2 + \kappa_3 R$, where $\kappa = (\kappa_0, \kappa_1, \kappa_2, \kappa_3)^T$ and $X_1^{ximpr} = RX_1 + (1 - R)E(X_1|X_2, R = 1)$. And build still working model for $E(X_1|X_2, R = 1)$ as $p(X_2, \eta) = \eta_0 + \eta_1^* X_2$. Then

$$\begin{aligned} \hat{\mu}_1^{ximpr} = & \frac{1}{n} \sum_{i=1}^n \frac{A_i Y_i}{\hat{\pi}} \\ & - (A_i - \hat{\pi}) \hat{\pi}^{-1} \left\{ (\kappa_0^* + \kappa_1^* \eta_0^*) + \kappa_1^* R_i X_{1i} + (\kappa_3 - \kappa_1^* \eta_0^*) R_i \right. \\ & \left. + (\eta_1^* \kappa_1^* + \kappa_2^*) X_{2i} - \eta_1^* \kappa_1^* R_i X_{2i} \right\}. \end{aligned} \quad (\text{A.3})$$

Thus, there is no restriction on the coefficient of R for the modified imputation method.

A.5 The asymptotic equivalence of ANCOVA1 and the augmented method

Based on results from Zhang, Zhang et al. (2008) for ANCOVA1: $\hat{\mu}_1^{ANCOVA1} = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{A_i Y_i}{\hat{\pi}} - (A_i - \hat{\pi}) \hat{\pi}^{-1} q^{ANCOVA1}(X_i) \right\}$ where $q^{ANCOVA1}(X_i) = \Sigma_{XY}^T \Sigma_{XX}^{-1} X_i$, $\Sigma_{XY} = E[(X - E(X))(Y - E(Y))]$ and $\Sigma_{XX} = E[(X - E(X))(X - E(X))]$. For the augmented method: $\hat{\mu}_1^{aug} = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{A_i Y_i}{\hat{\pi}} - (A_i - \hat{\pi}) \hat{\pi}^{-1} q^{aug}(X_i) \right\}$ where $q^{aug}(X_i) = \Sigma_{XY}^{T(1)} \Sigma_{XX}^{-1(1)} X_i$, $\Sigma_{XY}^{(1)} = E[(X - E(X))(Y - E(Y)) | A = 1]$ and $\Sigma_{XX}^{(1)} = E[(X - E(X))(X - E(X)) | A = 1]$. Therefore, if $Cov(X, Y)$ and $Var(X)$ are both independent of A , then the augmented method is asymptotically equivalent to ANCOVA1.

APPENDIX B

Appendix Chapter 3

B.1 Simulation 1: Missing Mechanisms

Table B.1: Missing mechanisms, percentages, and corresponding missing models.

Missing Mechanisms	Missing Percentage (%)	Missing Model
MCAR	20	$R \sim \text{Bernoulli}(0.2)$
	40	$R \sim \text{Bernoulli}(0.4)$
	70	$R \sim \text{Bernoulli}(0.7)$
MAR General	20	$\text{logit}(\Pr(R = 1)) = 0.2 + 0.3U$
	40	$\text{logit}(\Pr(R = 1)) = 0.16 + 0.02U$
	70	$\text{logit}(\Pr(R = 1)) = -1.1 + 0.02U$
MAR FIMAR	20	$\text{logit}(\Pr(R = 1)) = 0.9 + 1.6u_2 + X_2$
	40	$\text{logit}(\Pr(R = 1)) = 0.2 + 0.6u_2 + X_2$
	70	$\text{logit}(\Pr(R = 1)) = -1.3 + 0.6u_2 + X_2$
MNAR XZ	20	$\text{logit}(\Pr(R = 1)) = 2.1 + X_1 + X_2$
	40	$\text{logit}(\Pr(R = 1)) = 0.6 + X_1 + X_2$
	70	$\text{logit}(\Pr(R = 1)) = -1.3 + X_1 + X_2$
MNAR FIMAR	20	$\text{logit}(\Pr(R = 1)) = 0.56 + 1.8u_2$
	40	$\text{logit}(\Pr(R = 1)) = 0.8u_2$
	70	$\text{logit}(\Pr(R = 1)) = -1.25 + 0.8u_2$
MNAR CIMAR	20	$\text{logit}(\Pr(R = 1)) = 0.56 + 1.8u_1$
	40	$\text{logit}(\Pr(R = 1)) = 0.8u_1$
	70	$\text{logit}(\Pr(R = 1)) = -1.25 + 0.8u_1$

Notes: u_1 and u_2 are independent uniform random variables on $(0,1)$, used to generate the survival time T and censoring time C , respectively.

B.2 Simulation 2: Missing Mechanisms

Table B.2: Missing mechanisms, percentages, and corresponding missing models.

Missing Mechanisms	Missing Percentage (%)	Missing Model
MCAR	20	$R_1, \dots, R_8 \sim \text{Bernoulli}(0.2)$
	40	$R_1, \dots, R_8 \sim \text{Bernoulli}(0.4)$
	70	$R_1, \dots, R_8 \sim \text{Bernoulli}(0.7)$
MAR General	20	$\text{logit}(\Pr(R_1 = 1)) = 2.2 + Z_1 + \delta + 0.2U, \text{logit}(\Pr(R_2 = 1)) = 4.7 + Z_1 + Z_2$ $\text{logit}(\Pr(R_3 = 1)) = 2.8 + 0.2U, \text{logit}(\Pr(R_4 = 1)) = 1.8 + \delta + 0.3U$ $\text{logit}(\Pr(R_5 = 1)) = 2.2 + Z_1 + \delta + 0.2U, \text{logit}(\Pr(R_6 = 1)) = 4.7 + Z_1 + Z_2$ $\text{logit}(\Pr(R_7 = 1)) = 2.2 + Z_1 + \delta + 0.2U, \text{logit}(\Pr(R_8 = 1)) = 4.7 + Z_1 + Z_2$
	40	$\text{logit}(\Pr(R_1 = 1)) = 1.2 + Z_1 + \delta + 0.2U, \text{logit}(\Pr(R_2 = 1)) = 3.7 + Z_1 + Z_2$ $\text{logit}(\Pr(R_3 = 1)) = 1.8 + 0.2U, \text{logit}(\Pr(R_4 = 1)) = 0.8 + \delta + 0.3U$ $\text{logit}(\Pr(R_5 = 1)) = 1.2 + Z_1 + \delta + 0.2U, \text{logit}(\Pr(R_6 = 1)) = 3.7 + Z_1 + Z_2$ $\text{logit}(\Pr(R_7 = 1)) = 1.8 + 0.2U, \text{logit}(\Pr(R_8 = 1)) = 0.8 + \delta + 0.3U$
	70	$\text{logit}(\Pr(R_1 = 1)) = -0.1 + Z_1 + \delta + 0.2U, \text{logit}(\Pr(R_2 = 1)) = 2.4 + Z_1 + Z_2$ $\text{logit}(\Pr(R_3 = 1)) = 0.7 + 0.2U, \text{logit}(\Pr(R_4 = 1)) = -0.3 + \delta + 0.3U$ $\text{logit}(\Pr(R_5 = 1)) = -0.1 + Z_1 + \delta + 0.2U, \text{logit}(\Pr(R_6 = 1)) = 2.4 + Z_1 + Z_2$ $\text{logit}(\Pr(R_7 = 1)) = 0.7 + 0.2U, \text{logit}(\Pr(R_8 = 1)) = -0.3 + \delta + 0.3U$
MAR FIMAR	20	$\text{logit}(\Pr(R_1 = 1)) = 3.1 + u_2, \text{logit}(\Pr(R_2 = 1)) = 3.4 + Z_1 + u_2$ $\text{logit}(\Pr(R_3 = 1)) = 3.4 + Z_2 + u_2, \text{logit}(\Pr(R_4 = 1)) = 3.1 + u_2$ $\text{logit}(\Pr(R_5 = 1)) = 3.1 + u_2, \text{logit}(\Pr(R_6 = 1)) = 3.4 + Z_1 + u_2$ $\text{logit}(\Pr(R_7 = 1)) = 3.4 + Z_2 + u_2, \text{logit}(\Pr(R_8 = 1)) = 3.1 + u_2$
	40	$\text{logit}(\Pr(R_1 = 1)) = 2.2 + u_2, \text{logit}(\Pr(R_2 = 1)) = 2.6 + Z_1 + u_2$ $\text{logit}(\Pr(R_3 = 1)) = 2.6 + Z_2 + u_2, \text{logit}(\Pr(R_4 = 1)) = 2.2 + u_2$ $\text{logit}(\Pr(R_5 = 1)) = 2.2 + u_2, \text{logit}(\Pr(R_6 = 1)) = 2.6 + Z_1 + u_2$ $\text{logit}(\Pr(R_7 = 1)) = 2.6 + Z_2 + u_2, \text{logit}(\Pr(R_8 = 1)) = 2.2 + u_2$
	70	$\text{logit}(\Pr(R_1 = 1)) = 1.2 + u_2, \text{logit}(\Pr(R_2 = 1)) = 1.5 + Z_1 + u_2$ $\text{logit}(\Pr(R_3 = 1)) = 1.5 + Z_2 + u_2, \text{logit}(\Pr(R_4 = 1)) = 1.2 + u_2$ $\text{logit}(\Pr(R_5 = 1)) = 1.2 + u_2, \text{logit}(\Pr(R_6 = 1)) = 1.5 + Z_1 + u_2$ $\text{logit}(\Pr(R_7 = 1)) = 1.5 + Z_2 + u_2, \text{logit}(\Pr(R_8 = 1)) = 1.2 + u_2$

B.3 Simulation 2: Missing Mechanisms Continued

Table B.3: Missing mechanisms, percentages, and corresponding missing models.

Missing Mechanisms	Missing Percentage (%)	Missing Model
MNAR XZ	20	$\text{logit}(\Pr(R_1 = 1)) = 4.4 + X_1 + Z_1, \text{logit}(\Pr(R_2 = 1)) = 4.4 + X_2 + Z_2$
		$\text{logit}(\Pr(R_3 = 1)) = 5.3 + X_3 + Z_1 + Z_2, \text{logit}(\Pr(R_4 = 1)) = 5.3 + X_2 + X_4 + Z_1$
		$\text{logit}(\Pr(R_5 = 1)) = 4.4 + X_5 + Z_1, \text{logit}(\Pr(R_6 = 1)) = 4.4 + X_6 + Z_2$
		$\text{logit}(\Pr(R_7 = 1)) = 5.3 + X_7 + Z_1 + Z_2, \text{logit}(\Pr(R_8 = 1)) = 5.3 + X_8 + X_6 + Z_1$
	40	$\text{logit}(\Pr(R_1 = 1)) = 3.1 + X_1 + Z_1, \text{logit}(\Pr(R_2 = 1)) = 3.1 + X_2 + Z_2$
		$\text{logit}(\Pr(R_3 = 1)) = 3.8 + X_3 + Z_1 + Z_2, \text{logit}(\Pr(R_4 = 1)) = 3.8 + X_2 + X_4 + Z_1$
		$\text{logit}(\Pr(R_5 = 1)) = 3.1 + X_5 + Z_1, \text{logit}(\Pr(R_6 = 1)) = 3.1 + X_6 + Z_2$
		$\text{logit}(\Pr(R_7 = 1)) = 3.8 + X_7 + Z_1 + Z_2, \text{logit}(\Pr(R_8 = 1)) = 3.8 + X_8 + X_6 + Z_1$
	70	$\text{logit}(\Pr(R_1 = 1)) = 1.5 + X_1 + Z_1, \text{logit}(\Pr(R_2 = 1)) = 1.5 + X_2 + Z_2$
		$\text{logit}(\Pr(R_3 = 1)) = 1.9 + X_3 + Z_1 + Z_2, \text{logit}(\Pr(R_4 = 1)) = 1.9 + X_2 + X_4 + Z_1$
		$\text{logit}(\Pr(R_5 = 1)) = 1.5 + X_5 + Z_1, \text{logit}(\Pr(R_6 = 1)) = 1.5 + X_6 + Z_2$
		$\text{logit}(\Pr(R_7 = 1)) = 1.9 + X_7 + Z_1 + Z_2, \text{logit}(\Pr(R_8 = 1)) = 1.9 + X_8 + X_6 + Z_1$
MNAR FIMAR	20	$\text{logit}(\Pr(R_1 = 1)) = 3.1 + u_2, \text{logit}(\Pr(R_2 = 1)) = 3.4 + Z_1 + u_2$
		$\text{logit}(\Pr(R_3 = 1)) = 3.4 + Z_2 + u_2, \text{logit}(\Pr(R_4 = 1)) = 3.1 + u_2$
		$\text{logit}(\Pr(R_5 = 1)) = 3.1 + u_2, \text{logit}(\Pr(R_6 = 1)) = 3.4 + Z_1 + u_2$
		$\text{logit}(\Pr(R_7 = 1)) = 3.4 + Z_2 + u_2, \text{logit}(\Pr(R_8 = 1)) = 3.1 + u_2$
	40	$\text{logit}(\Pr(R_1 = 1)) = 2.2 + u_2, \text{logit}(\Pr(R_2 = 1)) = 2.6 + Z_1 + u_2$
		$\text{logit}(\Pr(R_3 = 1)) = 2.6 + Z_2 + u_2, \text{logit}(\Pr(R_4 = 1)) = 2.2 + u_2$
		$\text{logit}(\Pr(R_5 = 1)) = 2.2 + u_2, \text{logit}(\Pr(R_6 = 1)) = 2.6 + Z_1 + u_2$
		$\text{logit}(\Pr(R_7 = 1)) = 2.6 + Z_2 + u_2, \text{logit}(\Pr(R_8 = 1)) = 2.2 + u_2$
	70	$\text{logit}(\Pr(R_1 = 1)) = 1.2 + u_2, \text{logit}(\Pr(R_2 = 1)) = 1.5 + Z_1 + u_2$
		$\text{logit}(\Pr(R_3 = 1)) = 1.5 + Z_2 + u_2, \text{logit}(\Pr(R_4 = 1)) = 1.2 + u_2$
		$\text{logit}(\Pr(R_5 = 1)) = 1.2 + u_2, \text{logit}(\Pr(R_6 = 1)) = 1.5 + Z_1 + u_2$
		$\text{logit}(\Pr(R_7 = 1)) = 1.5 + Z_2 + u_2, \text{logit}(\Pr(R_8 = 1)) = 1.2 + u_2$
MNAR CIMAR	20	$\text{logit}(\Pr(R_1 = 1)) = 3.1 + u_1, \text{logit}(\Pr(R_2 = 1)) = 3.4 + Z_1 + u_1$
		$\text{logit}(\Pr(R_3 = 1)) = 3.4 + Z_2 + u_1, \text{logit}(\Pr(R_4 = 1)) = 3.1 + u_1$
		$\text{logit}(\Pr(R_5 = 1)) = 3.1 + u_1, \text{logit}(\Pr(R_6 = 1)) = 3.4 + Z_1 + u_1$
		$\text{logit}(\Pr(R_7 = 1)) = 3.4 + Z_2 + u_1, \text{logit}(\Pr(R_8 = 1)) = 3.1 + u_1$
	40	$\text{logit}(\Pr(R_1 = 1)) = 2.2 + u_1, \text{logit}(\Pr(R_2 = 1)) = 2.6 + Z_1 + u_1$
		$\text{logit}(\Pr(R_3 = 1)) = 2.6 + Z_2 + u_1, \text{logit}(\Pr(R_4 = 1)) = 2.2 + u_1$
		$\text{logit}(\Pr(R_5 = 1)) = 2.2 + u_1, \text{logit}(\Pr(R_6 = 1)) = 2.6 + Z_1 + u_1$
		$\text{logit}(\Pr(R_7 = 1)) = 2.6 + Z_2 + u_1, \text{logit}(\Pr(R_8 = 1)) = 2.2 + u_1$
	70	$\text{logit}(\Pr(R_1 = 1)) = 1.2 + u_1, \text{logit}(\Pr(R_2 = 1)) = 1.5 + Z_1 + u_1$
		$\text{logit}(\Pr(R_3 = 1)) = 1.5 + Z_2 + u_1, \text{logit}(\Pr(R_4 = 1)) = 1.2 + u_1$
		$\text{logit}(\Pr(R_5 = 1)) = 1.2 + u_1, \text{logit}(\Pr(R_6 = 1)) = 1.5 + Z_1 + u_1$
		$\text{logit}(\Pr(R_7 = 1)) = 1.5 + Z_2 + u_1, \text{logit}(\Pr(R_8 = 1)) = 1.2 + u_1$

B.4 Figures

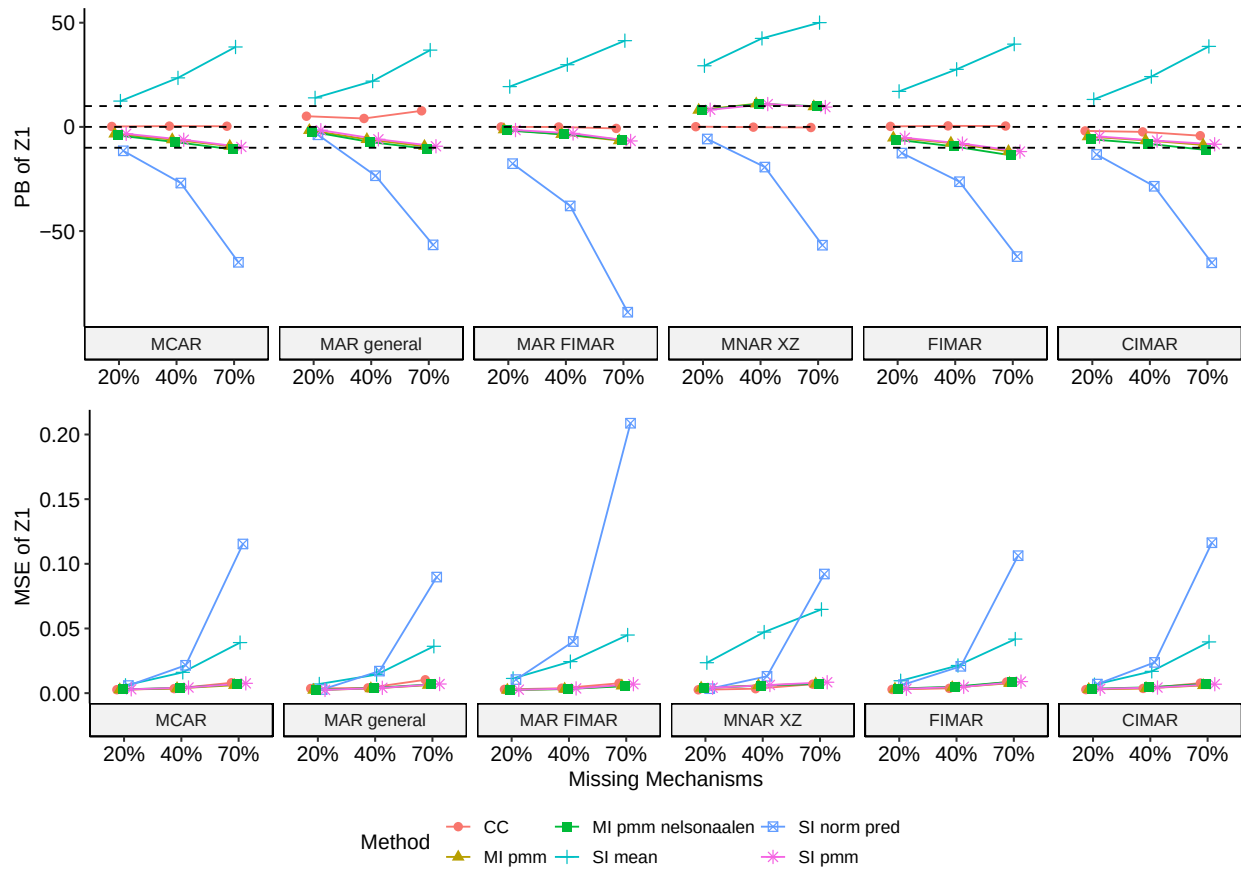


Figure B.1: PB and MC MSE of Z_1 in simulation 1. Dashed lines represent the $\pm 10\%$ PB.

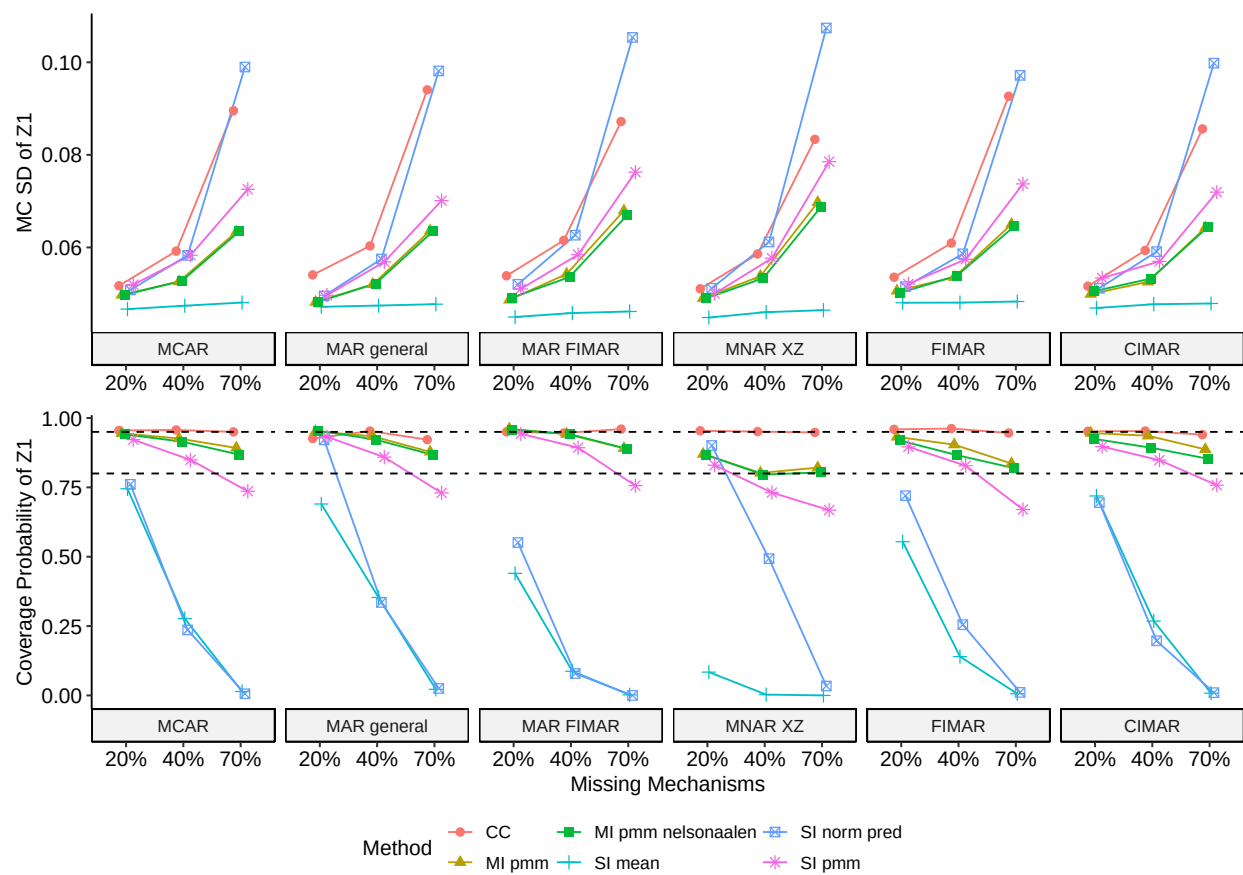


Figure B.2: SD and Cov. prob. of Z_1 in simulation 1. Dashed lines represent the 95% and 80% coverage probability.

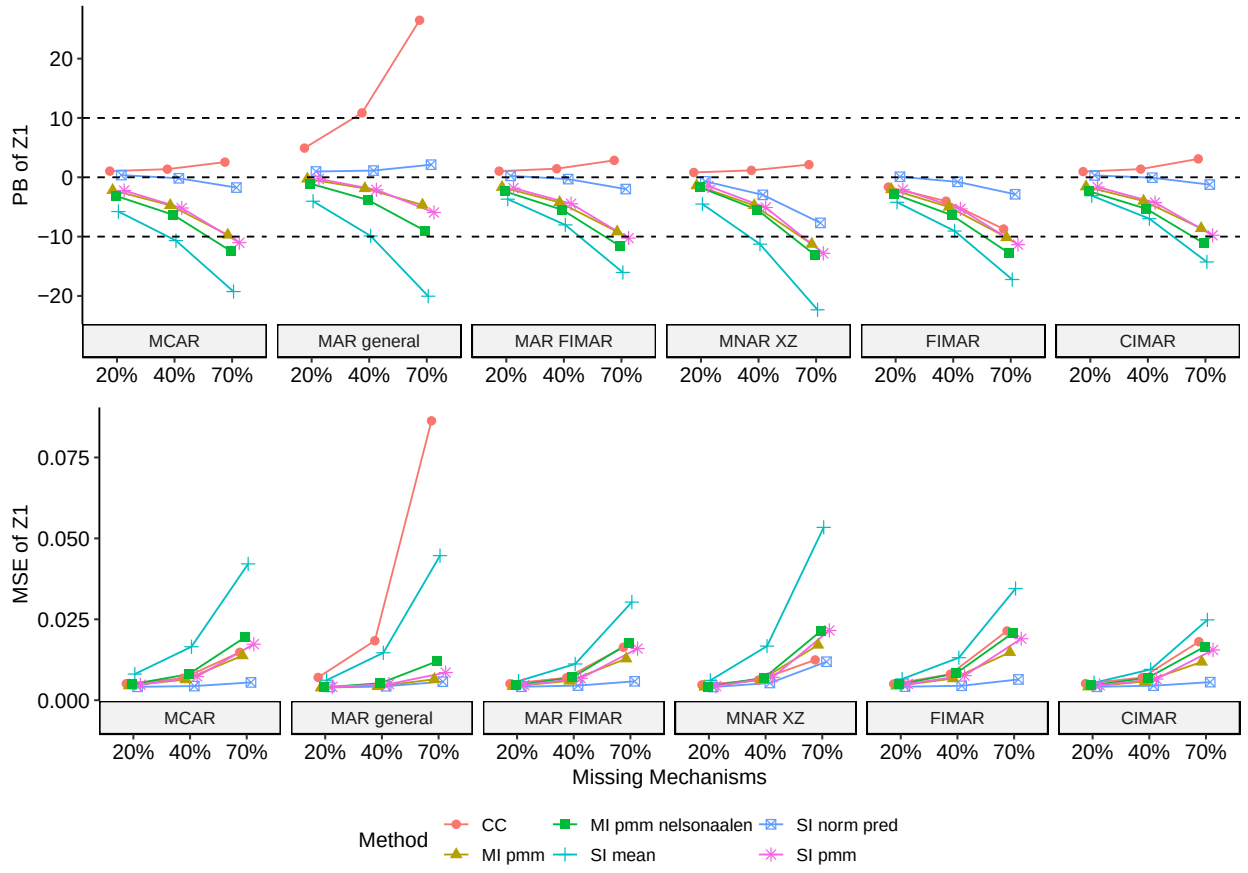


Figure B.3: PB and MC MSE of Z_1 in simulation 2. Dashed lines represent the $\pm 10\%$ PB.

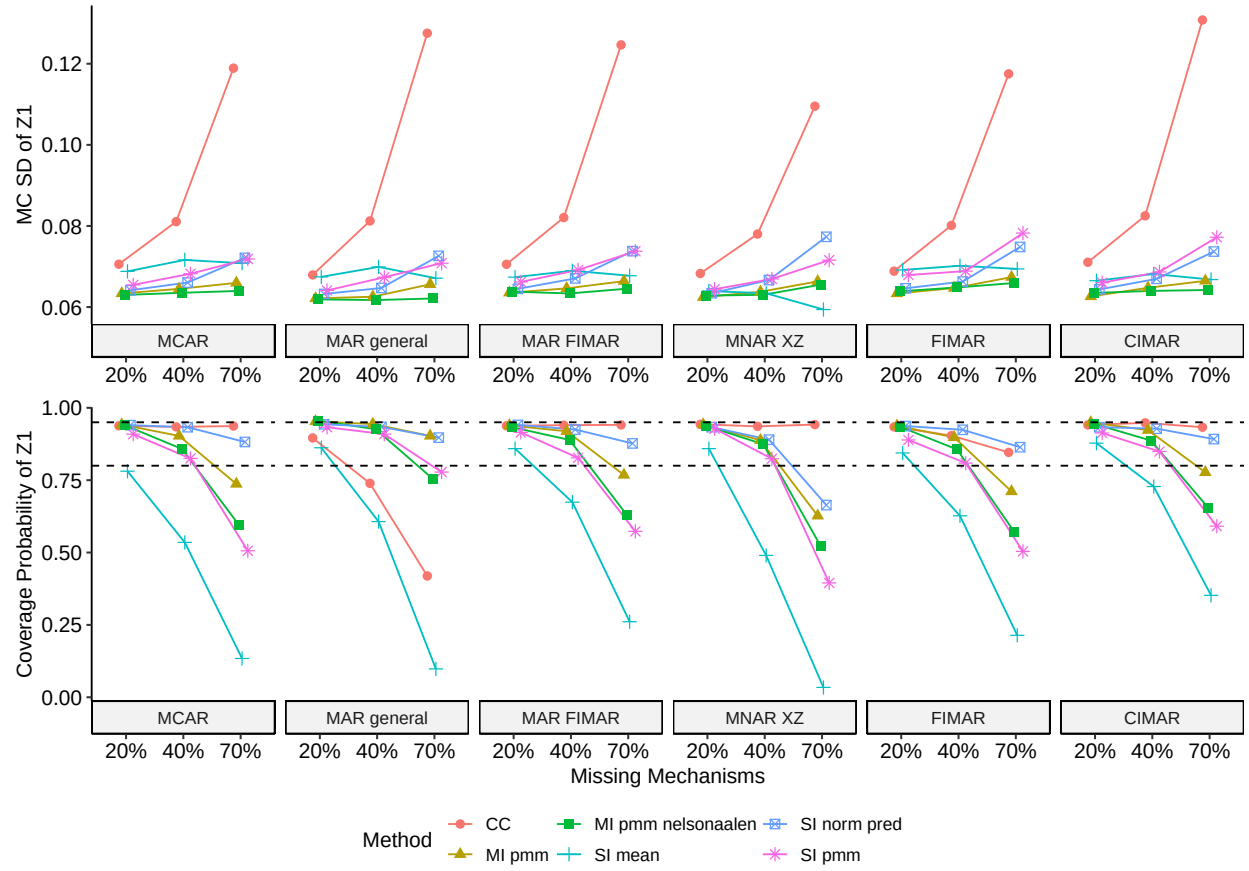


Figure B.4: SD and Cov. prob. of Z_1 in simulation 2. Dashed lines represent the 95% and 80% coverage probability.

APPENDIX C

Appendix Chapter 4

C.1 Joint survival function in toy example and its corresponding copula

$Y \sim \exp(\lambda)$, $X \mid Y = y \sim \text{uniform}(0, y)$. Therefore,

$$f_Y(y) = \lambda \exp(-\lambda y) I(y > 0), \quad f_{X|Y}(x \mid y) = \frac{1}{y} I(0 < x < y),$$

$$f_{X,Y}(x, y) = \lambda \exp(-\lambda y) \frac{1}{y} I(0 < x < y),$$

$$f_X(x) = \int_x^\infty f_{X,Y}(x, y) dy = \int_x^\infty \lambda \exp(-\lambda y) \frac{1}{y} I(x > 0) dy.$$

The joint survival function is given by:

$$S_{X,Y}(x, y) = \begin{cases} S_X(x), & \text{if } x \geq y, \\ (y - x)f_X(y) + S_X(y), & \text{if } x < y. \end{cases}$$

Substituting $u = S_X(x)$ and $v = S_Y(y)$, we have:

$$\begin{aligned} C(u, v) &= S(S_X^{-1}(u), S_Y^{-1}(v)) \\ &= \begin{cases} u, & \text{if } u < S_X \circ S_Y^{-1}(v), \\ (S_Y^{-1}(v) - S_X^{-1}(u)) f_X \circ S_Y^{-1}(v) + S_X \circ S_Y^{-1}(v), & \text{if } u \geq S_X \circ S_Y^{-1}(v). \end{cases} \end{aligned}$$

C.2 Model restrictions to satisfy the order restriction of cancer diagnosis and cancer death

Based on Equation (4.1), We have marginal survival functions $S_{T_k^*}, k = 1, 2$. We also have the known copula (4.2). Then if we plug in $u = S_{T_1^*}(t_1), v = S_{T_2^*}(t_2)$. it's easy to derive the pdf of T_1^* and T_2^* can be expressed as:

$$f_{T_1^*, T_2^*}(t_1, t_2) = \begin{cases} \frac{\partial Q(t_1)}{\partial t_1} - f'_X(P(t_2)) \frac{\partial P(t_2)}{\partial t_2}, & \text{when } t_1 \leq M(t_2), \\ 0, & \text{when } t_1 > M(t_2). \end{cases}$$

The order restriction $T_1^* < T_2^*$ with probability 1 can be expressed as $\Pr(T_1^* < T_2^*) = 1$.

1. When $M'(t_2) < 1$:

$$\begin{aligned} \Pr(T_1^* < T_2^*) &= \int_0^\infty \int_0^{M(t_2)} f_{T_1^*, T_2^*}(t_1, t_2) dt_1 dt_2 \\ &= \int_0^\infty Q(a) - f'_X(P(t_2)) \frac{\partial P(t_2)}{\partial t_2} \Big|_{a=0}^{M(t_2)} dt_2 \\ &= - \int_0^\infty P(t_2) f'_X(P(t_2)) \frac{\partial P(t_2)}{\partial t_2} dt_2 \\ &= - \int_0^\infty P(t_2) df_X(P(t_2)) \\ &= - \left[P(t_2) f_X(P(t_2)) \Big|_0^\infty - F_X(P(t_2)) \Big|_0^\infty \right] \\ &= 1. \end{aligned}$$

The first term is 0 because $f_X(x)$ can be approximated by $-x \ln x - \gamma x + x^2 - \frac{x^3}{4} + \dots$ as $x \rightarrow 0^+$, and by $\frac{e^{-x}}{x} (1 - \frac{1}{x} + \frac{1}{x^2} + \dots)$ as $x \rightarrow \infty$.

2. When $M'(t_2) > 1$:

$$\begin{aligned} \Pr(T_1^* < T_2^*) &= - \int_0^\infty Q(t_2) df_X(P(t_2)) \\ &= - \left[Q(t_2) f_X(P(t_2)) \Big|_0^\infty - \int_0^\infty f_X(P(t_2)) dQ(t_2) \right] \\ &= - \left[\frac{Q(t_2)}{P(t_2)} - P(t_2) f_X(P(t_2)) \Big|_0^\infty - \int_0^\infty \frac{f_X(P(t_2))}{f_X \circ Q(t_2)} f_{T_1^*}(t_2) dt_2 \right]. \end{aligned}$$

If $P(t_2) = Q(t_2)$ for all $t_2 \in (0, \infty)$, then $\Pr(T_1^* < T_2^*) = 1$. That is to say, when $M'(t_2) < 1$, we need $P(t_2) = Q(t_2)$ for all $t_2 \in (0, \infty)$ to ensure $\Pr(T_1^* < T_2^*) = 1$.

That is:

$$S_{T_1^*}(t_2) = S_X \circ S_Y^{-1} \circ S_{T_2^*}(t_2).$$

3. When $M'(t_2) = 1$: This is a trivial case where $S_{T_1^*, T_2^*}(t_1, t_2) = S_{T_1^*}(t_1)$.

C.3 No lambda information contained in data

We use the inverse transformation method to generate the survival times for cancer diagnosis and cancer death. As an example, consider the time to cancer diagnosis:

$$S_{T_1^*}(t) = u = S_X \left[\frac{\eta_2 H_2(t)}{\lambda} \right],$$

where u follows a uniform distribution on $(0, 1)$. Let

$$\frac{\eta_2 H_2(t)}{\lambda} = x.$$

Then,

$$S_X(x) = u = e^{-\lambda x} - \lambda x E_1(\lambda x).$$

If $z = \lambda x = \eta_2 H_2(t)$, we have:

$$S_X(z) = e^{-z} - z E_1(z) = u.$$

Thus, the inverse of S_X does not contain information about λ , and consequently, neither does the inverse of the marginal survival function of T_1^* .

C.4 EM algorithm

In our model, when cancer diagnosis or cancer death is censored, the time-to-diagnosis T_1^* or time-to-death T_2^* is missing, and the EM algorithm can be employed to estimate the baseline hazards (Tsodikov, 2003).

The conditional (complete data) score function can be expressed as:

$$U_0(s) = \frac{\partial \log L_{\text{com}}}{\partial dH_2(s)}$$

Here, L_{com} represents the complete data likelihood, assuming both the time-to-diagnosis and the time-to-death are observed.

In the $(k + 1)$ th iteration of the EM algorithm, the E-step computes the unconditional score function, defined as:

$$U(s) = \mathbb{E}[U_0(s)],$$

where the expectation is taken with respect to the distribution of the missing data given the observed data and the k th iteration estimate of dH_2 . In the M-step, we solve $U(s) = 0$ to obtain the updated estimate $dH_2^{(k+1)}(s)$.

The remainder of this section is organized as follows. First, we derive the E-step for $\{dH_2\}$ under three distinct scenarios: both events observed, diagnosis observed but death censored, and neither event observed. Next, we derive the M-step to maximize the likelihood with respect to $\{dH_2\}$, which results in a closed-form expression analogous to the weighted Breslow-type estimators proposed by Chen (2012). The following notations are introduced:

$$\begin{aligned} G_1(a, b) &= \int_a^b G_{1.t2m} dH_2(m), \\ G_{1.t2m} &= \frac{\eta_2^2 \left[1 + \frac{1}{\lambda P(t_1)}\right] [-f'_X(P(m))]}{\lambda f_X \circ P(t_2)}, \\ G_{2.t2s} &= \frac{[-f'_X(P(s))]\eta_2}{\lambda f_X \circ P(t_2)}, \\ J_1(a, b) &= \int_a^b J_{1.tm} dH_2(m), \\ J_{1.tm} &= \frac{\left(1 + \frac{1}{\lambda P(t_1)}\right) \eta_2^2 [P(m) - P(t_1)] [-f'_X(P(m))]}{S_X \circ P(t_1) \lambda}, \\ J_{2.ts} &= \frac{f_X \circ P(s) \eta_2}{S_X \circ P(t_1) \lambda}, \\ J_{3.ts} &= \frac{[P(s) - P(t_1)] [-f'_X(P(s))]\eta_2}{S_X \circ P(t_1) \lambda}. \end{aligned}$$

C.4.1 E step

1. Subject has a diagnosis at T_1 and death at T_2 ($\delta_1 = \delta_2 = 1$):

$$\begin{aligned} L_{\text{com}} &= \frac{\partial P(t_1)}{\partial t_1} (-f'_X(P(t_2))) \frac{\partial P(t_2)}{\partial t_2}. \\ U_{0,dH_2(s)} &= \frac{\partial \log L_{\text{com}}}{\partial dH_2(s)} \\ &= \frac{I(t_1 = s)}{dH_2(s)} + \frac{I(t_2 = s)}{dH_2(s)} - \left[1 + \frac{1}{\lambda P(t_2)}\right] \eta_2 I(t_2 \geq s). \end{aligned}$$

Since no data is missing in this case:

$$\begin{aligned} U_{dH_2(s)} &= U_{0,dH_2(s)} \\ &= \frac{I(t_1 = s)}{dH_2(s)} + \frac{I(t_2 = s)}{dH_2(s)} - \left[1 + \frac{1}{\lambda P(t_2)}\right] \eta_2 I(t_2 \geq s). \end{aligned} \quad (C.1)$$

2. Subject has a diagnosis at T_1 and is censored at T_2 ($\delta_1 = 1, \delta_2 = 0, T_1^* = t_1, T_2^* = m$):
Here, for L_{com} , we assume $T_2^* = m$ is observed. Then:

$$U_{0,dH_2(s)} = \frac{I(t_1 = s)}{dH_2(s)} + \frac{I(m = s)}{dH_2(s)} - \left[1 + \frac{1}{\lambda P(m)}\right] \eta_2 I(m \geq s).$$

(a) When $s \leq t_2 < m$:

$$U_{0,dH_2(s)} = \frac{I(t_1 = s)}{dH_2(s)} - \left[1 + \frac{1}{\lambda P(m)}\right] \eta_2.$$

$$\begin{aligned} U_{dH_2(s)} &= \mathbb{E}[U_{0,dH_2(s)}] \\ &= \frac{I(t_1 = s)}{dH_2(s)} - \int_{t_2}^{\infty} \eta_2 \left(1 + \frac{1}{\lambda P(m)}\right) \frac{\Pr(T_1^* = t_1, T_2^* = m)}{\Pr(T_1^* = t_1, T_2^* > t_2)} dm \\ &= \frac{I(t_1 = s)}{dH_2(s)} - G_1(t_2, \infty). \end{aligned}$$

(b) When $s \geq t_2$:

$$U_{0,dH_2(s)} = \frac{I(m = s)}{dH_2(s)} - \left(1 + \frac{1}{\lambda P(m)}\right) \eta_2 I(m \geq s).$$

$$\begin{aligned} U_{dH_2(s)} &= \mathbb{E}[U_{0,dH_2(s)}] \\ &= \frac{1}{dH_2^{(k+1)}(s)} G_{2,t_2s} dH_2^{(k)}(s) - G_1(s, \infty). \end{aligned}$$

Combining results above: when $\Delta_1 = 1, \Delta_2 = 0$:

$$\begin{aligned} U_{dH_2(s)} &= Y_2(s) \left\{ \frac{I(t_1 = s)}{dH_2(s)} - G_1(t_1, \infty) \right\} \\ &\quad + [1 - Y_2(s)] \left\{ \frac{dH_2^{(k)}(s)}{dH_2^{(k+1)}(s)} G_{2,t_2s} - G_1(s, \infty) \right\}. \end{aligned} \quad (C.2)$$

3. Subject is censored before any event occurs ($\Delta_1 = 0, \Delta_2 = 0, T_1 = T_2, T_1^* = l, T_2^* =$

$m, t_1 = t_2 < l < m$):

$$U_{0,dH_2(s)} = \frac{I(l=s)}{dH_2(s)} + \frac{I(m=s)}{dH_2(s)} - \left(1 + \frac{1}{\lambda P(m)}\right) \eta_2 I(m \geq s).$$

(a) When $s \leq t_1 = t_2 < l < m$:

$$U_{0,dH_2(s)} = - \left(1 + \frac{1}{\lambda P(m)}\right) \eta_2 I(m \geq s).$$

$$\begin{aligned} U_{dH_2(s)} &= \mathbb{E}[U_{0,dH_2(s)}] \\ &= - \int_{t_2}^{\infty} \int_{t_1}^{\infty} \left[1 + \frac{1}{\lambda P(m)}\right] \eta_2 \frac{\Pr(T_1^* = l, T_2^* = m)}{\Pr(T_1^* > t_1, T_2^* > t_2)} dt_1 dm \\ &= -J_1(t_2, \infty). \end{aligned}$$

(b) When $s \geq t_1 = t_2$:

$$U_{0,dH_2(s)} = \frac{I(l=s)}{dH_2(s)} + \frac{I(m=s)}{dH_2(s)} - \left(1 + \frac{1}{\lambda P(m)}\right) \eta_2 I(m \geq s).$$

$$\begin{aligned} U_{dH_2(s)} &= \mathbb{E}[U_{0,dH_2(s)}] \\ &= \frac{dH_2^{(k)}(s)}{dH_2^{(k+1)}(s)} (J_{2.ts} + J_{3.ts}) - J_1(s, \infty). \end{aligned}$$

Combining results above: when $\Delta_1 = 0, \Delta_2 = 0$:

$$\begin{aligned} U_{dH_2(s)} &= Y_2(s) [-J_1(t_2, \infty)] \\ &\quad + [1 - Y_2(s)] \left[\frac{dH_2^{(k)}(s)}{dH_2^{(k+1)}(s)} (J_{2.ts} + J_{3.ts}) - J_1(s, \infty) \right]. \end{aligned} \tag{C.3}$$

Combining (C.1), (C.2), and (C.3), we have:

$$U_{dH_2^{(j+1)}(s)} = \sum_{i=1}^n \left\{ \frac{dN_{1i}(s) + dN_{2i}(s)}{dH_2^{(j+1)}(s)} - \Psi_i^{(j)}(s) + \left[\frac{dH_2^{(j)}(s)}{dH_2^{(j+1)}(s)} - 1 \right] \theta_i^{(j)}(s) \right\} = 0,$$

where

$$\begin{aligned}
\Psi^{(j)}(s) &= \Delta_1 \Delta_2 \left(1 + \frac{1}{\lambda P(t_1)} \right) \eta_2 Y_2(s) \\
&\quad + \Delta_1 (1 - \Delta_2) \left\{ Y_2(s) G_1(t_2, \infty) - [1 - Y_2(s)] [G_{2,t_2s} - G_1(s, \infty)] \right\} \\
&\quad + (1 - \Delta_1) (1 - \Delta_2) \left\{ Y_2(s) J_1(t_2, \infty) - [1 - Y_2(s)] [J_{2,ts} + J_{3,ts} - J_1(s, \infty)] \right\}, \\
\theta^{(j)}(s) &= \Delta_1 (1 - \Delta_2) [1 - Y_2(s)] G_{2,t_2s} + (1 - \Delta_1) (1 - \Delta_2) [1 - Y_2(s)] (J_{2,ts} + J_{3,ts}).
\end{aligned}$$

C.4.2 M step

In the M-step, we maximize the function from the E-step by solving $\sum_{i=1}^n U_{i,dH_2}(s) = 0$, which gives

$$dH_2^{(j+1)}(s) = \frac{\sum_i (dN_{1i}(s) + dN_{2i}(s)) + \left[\sum_i \theta_i^{(j)}(s) \right] dH_2^{(j)}(s)}{\sum_i \left[\Psi_i^{(j)}(s) + \theta_i^{(j)}(s) \right]}.$$

The above equation solves for $dH_2(s)$ iteratively until convergence, and the estimator at convergence, $\widehat{dH_2}(s)$, is a consistent estimator of $dH_2(s)$.

BIBLIOGRAPHY

- American Cancer Society (2020). Prostate cancer statistics. <https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html>.
- Andriole, G. L., Crawford, E. D., Grubb III, R. L., Buys, S. S., Chia, D., Church, T. R., Fouad, M. N., Gelmann, E. P., Kvale, P. A., Reding, D. J., et al. (2009). Mortality results from a randomized prostate-cancer screening trial. *New England journal of medicine*, 360(13):1310–1319.
- Andriole, G. L., Crawford, E. D., Grubb III, R. L., Buys, S. S., Chia, D., Church, T. R., Fouad, M. N., Isaacs, C., Kvale, P. A., Reding, D. J., et al. (2012). Prostate cancer screening in the randomized prostate, lung, colorectal, and ovarian cancer screening trial: mortality results after 13 years of follow-up. *Journal of the National Cancer Institute*, 104(2):125–132.
- Assmann, S. F., Pocock, S. J., Enos, L. E., and Kasten, L. E. (2000). Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet*, 355(9209):1064–1069.
- Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., and Initiative*, A. D. N. (2015). Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Statistical methods in medical research*, 24(4):462–487.
- Benkeser, D., Díaz, I., Luedtke, A., Segal, J., Scharfstein, D., and Rosenblum, M. (2021). Improving precision and power in randomized trials for covid-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. *Biometrics*, 77(4):1467–1481.
- Chang, C.-R., Song, Y., Li, F., and Wang, R. (2022). Covariate adjustment in randomized experiments with incomplete baseline data. *arXiv preprint arXiv:2207.07890*.
- Chen, H. Y. and Little, R. J. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, 94(447):896–908.
- Chen, Y.-H. (2012). Maximum likelihood analysis of semicompeting risks data with semi-parametric regression models. *Lifetime data analysis*, 18:36–57.
- Emura, T. and Chen, Y.-H. (2018). *Analysis of survival data with dependent censoring: copula-based approaches*, volume 450. Springer.
- Fine, J. P., Jiang, H., and Chappell, R. (2001). On semi-competing risks data. *Biometrika*, 88(4):907–919.

- Food, Administration, D., et al. (2020). Covid-19: developing drugs and biological products for treatment or prevention. *Guidance for Industry*.
- Groenwold, R. H., White, I. R., Donders, A. R. T., Carpenter, J. R., Altman, D. G., and Moons, K. G. (2012). Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *Cmaj*, 184(11):1265–1269.
- Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335(15):1081–1090.
- Hsu, C.-H. and Yu, M. (2019). Cox regression analysis with missing covariates via nonparametric multiple imputation. *Statistical methods in medical research*, 28(6):1676–1688.
- Hu, C. and Tsodikov, A. (2014). Semiparametric regression analysis for time-to-event marked endpoints in cancer studies. *Biostatistics*, 15(3):513–525.
- Huang, X. and Liu, L. (2007). A joint frailty model for survival and gap times between recurrent events. *Biometrics*, 63(2):389–397.
- Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346.
- Jiang, F., Tian, L., Fu, H., Hasegawa, T., and Wei, L. (2019). Robust alternatives to ancova for estimating the treatment effect via a randomized comparative study. *Journal of the American Statistical Association*, 114(528):1854–1864.
- Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association*, 91(433):222–230.
- Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*. John Wiley & Sons.
- Kayembe, M. T., Jolani, S., Tan, F. E., and van Breukelen, G. J. (2020). Imputation of missing covariate in randomized controlled trials with a continuous outcome: Scoping review and new results. *Pharmaceutical statistics*, 19(6):840–860.
- Leon, S., Tsiatis, A. A., and Davidian, M. (2003). Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics*, 59(4):1046–1055.
- Lin, D., Sun, W., and Ying, Z. (1999). Nonparametric estimation of the gap time distribution for serial events with censored data. *Biometrika*, 86(1):59–70.
- Lipsitz, S. R., Ibrahim, J. G., and Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association*, 94(448):1147–1160.

- Little, R. J. (1992). Regression with missing x's: a review. *Journal of the American statistical association*, 87(420):1227–1237.
- Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.
- Marshall, A., Altman, D. G., Royston, P., and Holder, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC medical research methodology*, 10(1):1–16.
- Moons, K. G., Donders, R. A., Stijnen, T., and Harrell Jr, F. E. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of clinical epidemiology*, 59(10):1092–1101.
- Moore, K. L. and van der Laan, M. J. (2009). Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Statistics in medicine*, 28(1):39–64.
- Murphy, S. A. and Van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465.
- Nelsen, R. B. (2006). An introduction to copulas.
- Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in medicine*, 21(19):2917–2930.
- Qi, L., Wang, Y.-F., and He, Y. (2010). A comparison of multiple imputation and fully augmented weighted estimators for cox regression with missing covariates. *Statistics in medicine*, 29(25):2592–2604.
- Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., Solenberger, P., et al. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96.
- Rathouz, P. J. (2007). Identifiability assumptions for missing covariate data in failure time regression models. *Biostatistics*, 8(2):345–356.
- Rice, J. D. and Tsodikov, A. (2017). Semiparametric time-to-event modeling in the presence of a latent progression event. *Biometrics*, 73(2):463–472.
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

- Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, volume 1, pages 20–34. American Statistical Association Alexandria, VA, USA.
- Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. hoboken.
- Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.
- Rubin, D. B. (2018). Multiple imputation. In *Flexible Imputation of Missing Data, Second Edition*, pages 29–62. Chapman and Hall/CRC.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.
- Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, 7(2):147.
- 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.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.
- Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in medicine*, 27(23):4658–4677.
- Tsodikov, A. (2003). Semiparametric models: a generalized self-consistency approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(3):759–774.
- Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3):219–242.
- Van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, 18(6):681–694.
- Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67.
- Wang, C. and Chen, H. Y. (2001). Augmented inverse probability weighted estimator for cox missing covariate regression. *Biometrics*, 57(2):414–419.
- Wang, W. (2021). *Semiparametric Joint Models to Assess the Effect of Timing of the Indication for Treatment on Survival with Application to Cancer Screening*. PhD thesis.
- White, I. R. and Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in medicine*, 29(28):2920–2931.

- White, I. R. and Royston, P. (2009). Imputing missing covariate values for the cox model. *Statistics in medicine*, 28(15):1982–1998.
- White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399.
- Xu, J., Kalbfleisch, J. D., and Tai, B. (2010). Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics*, 66(3):716–725.
- Yang, L. and Tsiatis, A. A. (2001). Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321.
- Zhang, M., Tsiatis, A. A., and Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707–715.
- Zhao, A. and Ding, P. (2022). To adjust or not to adjust? estimating the average treatment effect in randomized experiments with missing covariates. *Journal of the American Statistical Association*, pages 1–11.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.