# IS COMPLETE CASE ANALYSIS APPROPRIATE FOR COX REGRESSION WITH MISSING COVARIATE DATA?

by

Min Zhu

---



A Thesis Submitted to the Faculty of the

MEL AND ENID ZUCKERMAN COLLEGE OF PUBLIC HEALTH

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

WITH A MAJOR IN BIOSTATISTICS

In the Graduate College

THE UNIVERSITY OF ARIZONA

2018

STATEMENT BY AUTHOR

The thesis titled ***Is Complete Case Analysis Appropriate for Cox Regression with Missing Covariate Data?*** prepared by ***Min Zhu*** has been submitted in partial fulfillment of requirements for a master's degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Min Zhu

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

April 4, 2018

Chiu-Hsieh (Paul) Hsu, PhD, Professor of Biostatistics — Date

**Acknowledgements**

I would first like to thank my thesis director and committee chair Dr. Chiu-Hsieh (Paul) Hsu of the department of Biostatistics in the Mel and Enid Zuckerman College of Public Health at the University of Arizona. Dr. Hsu's guidance has been instrumental at every step of this project. I could not ask more from a teacher or mentor.

I would also like to thank my committee members Dr. Melanie Bell and Dr. Dean Billheimer of the department of Biostatistics in the Mel and Enid Zuckerman College of Public Health at the University of Arizona. Dr. Bell and Dr. Billheimer provided advice and insights that were essential for completion of this project.

Finally, I would like to thank my family, friends, classmates, as well as the online programming and statistical community for all their help and support throughout the years.

**Table of Contents**

## Abstract

*Purpose:* Complete case analysis of survival datasets with missing covariates in Cox proportional hazards model relies heavily on strong and usually unverifiable missing mechanism assumptions such as missing completely at random (MCAR) to produce reasonable parameter estimates. Based on the nature of survival data, missing at random (MAR) for missing covariates can be further decomposed into 1) censoring ignorable missing at random (CIMAR) and 2) failure ignorable missing at random (FIMAR). Unlike MCAR and MAR, there are procedures to assess whether missingness of covariates in survival data are consistent with CIMAR or FIMAR. In my thesis, I investigate the performances of the complete case analysis under various missing mechanisms in Cox model and demonstrate the procedures for checking consistency with CIMAR or FIMAR.

*Experimental design:* For research involving missing data, simulation studies are especially useful while studying the performance of some estimation (e.g. complete case analysis) as all parameters are pre-specified and known. I simulate survival data with missing covariates under various missing data mechanisms including MCAR, missing at random (MAR), missing not at random (MNAR), CIMAR and FIMAR. I then perform complete case Cox regression on simulated datasets and compare results to determine which missingness mechanisms produce reasonable parameter estimates. Finally, I perform a two-step procedure to check whether covariate missingness is consistent with CIMAR or FIMAR on a real dataset as outlined by Rathouz (2006).

*Results:* This simulation study illustrates that when covariate missingness is FIMAR but not CIMAR, complete case Cox regression produces reasonable parameter estimates similar to when missingness is MCAR. When covariate missingness is CIMAR, complete case Cox regression

produces biased parameter estimates. The two-step procedure suggests covariate missingness in the Stanford heart transplant data is consistent with FIMAR.

*Conclusions:* Survival data with missing covariates that are FIMAR are appropriate for complete case analysis in Cox models. Survival data with missing covariates that are CIMAR are not appropriate for complete case analysis in Cox models. Under independent censoring, it should be possible for researchers to check the consistency of missing covariates in survival data with FIMAR and CIMAR assumptions. If missingness is consistent with FIMAR, complete case Cox regression should produce reasonable estimates. If missingness is consistent with CIMAR or if the data is inconsistent with both CIMAR and FIMAR, complete case Cox regression may produce biased estimates and researchers should consider sensitivity analyses.

## 1. Introduction

In survival analysis, Cox proportional hazards model is the most popular regression model for studying the relationships between predictors and survival time.[1] Unfortunately, survival studies are prone to missing covariate data due to various reasons. Missing data cannot be ignored as they may bias analyses results depending on the missing mechanism. In addition, how researchers choose to handle missing data will also affect the magnitude of bias, especially when missingness is not ignorable.[2] In Cox models, complete case analysis (CCA) is the most popular method for handling missing covariates and is also the default method for analyses in many statistical programs.[3] While popular and widely used in Cox regression, CCA relies on strong and often unjustifiable assumptions of the underlying missing mechanism to yield reasonable parameter estimates.[4]

It has been shown that when missingness is missing completely at random (MCAR), complete case Cox regression yields consistent estimates of the regression coefficients.[5] However, when missingness is missing at random (MAR), complete case Cox regression may produce biased estimates.[6] Based on the nature of survival data, Rathouz (2006) decomposed MAR for missing covariates into 1) censoring ignorable missing at random (CIMAR) and 2) failure ignorable missing at random (FIMAR). In some sense, CIMAR and FIMAR assumptions are weaker than MAR since they do not require conditioning on the entire observed data. In addition, under independent censoring one can use CIMAR or FIMAR as working assumption to check whether missing covariate data in survival analyses are consistent with CIMAR or FIMAR through a two-step procedure as outlined by Rathouz.[4] This makes CIMAR or FIMAR assumptions appealing for researchers since there are no such procedures to check whether missing covariate data in survival analyses are consistent with MCAR or MAR assumptions. The performance of CCA in Cox

models with missing covariates under FIMAR or CIMAR has not been thoroughly investigated. Also, the two-step procedure for checking consistency with FIMAR or CIMAR missing covariates is not well recognized. Therefore, in my thesis, I will first evaluate the performances of CCA in Cox models with missing covariate data under various missing mechanisms, including FIMAR and CIMAR, and then demonstrate the two-step procedure on a real dataset.

My thesis is organized as follows: In Section 2, I review Cox (proportional hazards) model. In Section 3, I review missing mechanisms in survival analysis with respect to covariates. In Section 4, I describe the two-step procedure to check survival datasets with missing covariates for consistency with CIMAR and FIMAR assumptions. In Section 5, I perform simulations to evaluate CCA in Cox models with missing covariate data. In Section 6, I demonstrate the two-step procedure on the Stanford heart transplant dataset. Discussion follows in Section 7.

## 2. Review of Cox Model

In this section, I begin by describing Cox proportional hazards model introduced by D.R. Cox in 1972.[7] It has been extensively studied and its theoretical properties have been well established.[8] Let *T* denote the failure time, *C* denote the censoring time, *Y=min(T,C)* denote the observed time and *D=I[T ≤ C]* denote the censoring indicator. Cox proportional hazards model assumes *T* has a hazard function of $\lambda(t)=\lambda_0(t)e^{\beta_x X}$, where $\lambda_0(t)$ is an unspecified baseline hazard function and X is a predictor. Based on the expression, it is a semi-parametric model since the baseline hazard function is unspecified. Also, it is a multiplicative hazards model and the hazard ratio (e.g. $e^{\beta_x}$) can be regarded as a relative risk measure for X.[9] The hazard ratio (HR) indicates whether the predictor is associated with increasing or decreasing risk by comparing with 1 and can be easily interpreted. These properties (semi-parametric and natural presentation of hazard ratio) make Cox model popular and widely used in survival analysis.

Often the estimation of the regression coefficient in Cox model is based on partial likelihood, which eliminates the nuisance parameter (baseline hazard). Let $r_i(\boldsymbol{\beta},t)= e^{\beta_x X_i} \equiv r_i^{(0)}(\boldsymbol{\beta},t)$ and $r_i^{(1)}=X_i\ r_i(\boldsymbol{\beta},t)$; i is the subject index. The partial likelihood estimator involves solving the following estimation equation:

$$U = \sum_{i=1}^{n} \left[ D_i \left\{ X_i - \frac{S^{(1)}(\boldsymbol{\beta}, T_i)}{S^{(0)}(\boldsymbol{\beta}, T_i)} \right\} \right] = 0,$$

where $S^{(m)}(\boldsymbol{\beta}, T_i) = n^{-1} \sum_{j=1}^{n} I(T_j \geq T_i) r_j^{(m)}(\boldsymbol{\beta}, T_i)$ for m=0, 1. It is easy to implement the estimation. When there is no missing data in X and censoring time is independent of failure time (i.e. independent censoring), the above partial likelihood estimator is consistent. When X is subject

to missing, the complete case Cox model based on the partial likelihood approach involves solving the following estimation equation:

$$U_{CC} = \sum_{i=1}^{n} \left[ D_i R_i \left\{ X_i - \frac{S_{cc}^{(1)}(\boldsymbol{\beta}, T_i)}{S_{cc}^{(0)}(\boldsymbol{\beta}, T_i)} \right\} \right] = 0,$$

where R is the missing indicator for X (i.e. R = 1 if X is observed; otherwise, 0) and $S_{cc}^{(m)}(\boldsymbol{\beta}, T_i) = n^{-1} \sum_{j=1}^{n} R_i I(T_j \geq T_i) r_j^{(m)}(\boldsymbol{\beta}, T_i)$ for m=0, 1. It is also easy to implement the CCA and it is consistent when the missingness is ignorable and censoring is independent of failure time. However, it loses efficiency due to discarding data from incomplete observations, especially when the missing rate is greater than 25%,[10] and is inconsistent when missingness depends on T or D.[6] The performance of the CCA highly depends on the underlying missing mechanism. Therefore, I review the missing mechanism for missing covariates in survival analysis in the next section.

## 3. Missing mechanism for missing covariates in survival analysis

In the missing data literature for regression with missing covariates, the missing mechanism can be classified into three categories: missing complete at random (MCAR), missing at random (MAR) and missing not at random (MNAR).[11] They are defined below.

### *3.1 MCAR*

When data is MCAR, it simply means the missingness does not depend on the missing covariate X:

$$R \perp\!\!\!\perp \mathrm{X},$$

where R is the missing indicator. In a regression setting with other fully observed covariates Z, MCAR can be expressed as the missingness does not depend on the missing covariate X or the outcome of interest (i.e. Y and D in survival analysis) conditional on the fully observed covariates Z:

$$R \perp\!\!\!\perp (\mathrm{Y}, \mathrm{D}, \mathrm{X}) | \mathrm{Z},$$

where R is the missing indicator, Y is the observed time and D is the censoring indicator.[4] A practical example of MCAR may be a study that stops a test which measures a specific covariate due to budget cuts. Assuming the tests were administered at random, missingness of the covariate is MCAR. Under MCAR, CCA is expected to generate consistent estimates of the regression coefficients in Cox model but will lose efficiency in estimation.

*3.2 MAR*

When data is MAR, the missingness does not depend on the missing covariate conditional on the observed data (i.e. Y, D and Z):[4]

$$R \perp\!\!\!\perp \mathrm{X} | (\mathrm{Y}, \mathrm{D}, \mathrm{Z}).$$

A practical example of MAR may be a case-cohort study using frozen tissue samples. Suppose an assay for a specific covariate was only performed at random on every other sample from subjects in the cohort that had an event. Missingness of the covariate is MAR. The CCA is expected to produce biased estimates of the regression coefficients in Cox model. The magnitude of bias depends on the relationship between the missingness and the observed data as well as the missing rate. Based on the nature of survival data, Rathouz decomposed MAR in survival analysis with missing covariates into two MAR-like missing mechanisms: censoring time ignorable missing at random (CIMAR) and failure time ignorable missing at random (FIMAR).

*3.3 CIMAR*

When data is CIMAR, the missingness does not depend the missing covariate X or censoring time conditional on the fully observed covariate Z and failure time T:

$$R \perp\!\!\!\perp (\mathrm{C}, \mathrm{X}) | (\mathrm{T}, \mathrm{Z}).$$

A practical example of CIMAR may be a study in which patients enroll over some years and censoring for all patients occurs on the same calendar date. Assume the study protocol changes over the course of the study and some previous tests (covariates) are no longer needed or protocol dictates new tests. In this example, the missingness of a covariate predicts the censoring time.[4]

*3.4 FIMAR*

When data is FIMAR, the missingness does not depend on the missing covariate X or failure time conditional on the fully observed covariates Z and censoring time:

$$R \perp\!\!\!\perp (\mathrm{T}, \mathrm{X}) | (\mathrm{C}, \mathrm{Z}).$$

A practical example of FIMAR may be expensive or invasive tests to measure a specific covariate. Suppose researchers administer such a test for patients who are more likely to test positive. The missingness of the covariate depends on the severity of disease. Thus the presence of the test and covariate predicts failure times.[4]

*3.5 MNAR*

When data is MNAR, the missingness depends on the missing covariate X itself:

$$R | (\mathrm{X}, \mathrm{Z}).$$

A practical example of MNAR might be a study of income and disease. Suppose income is self-reported and low-income subjects are less likely to report their incomes. Missingness of income is MNAR. Note that in survival analysis, under MNAR, if the missingness only depends on covariates (observed or unobserved) but not on failure time or censoring time, complete case Cox regression will produce reasonable parameter estimates.[4] However, if the MNAR missingness also depends on the outcome of interest,

$$R | (\mathrm{X}, \mathrm{Y}, \mathrm{D}, \mathrm{Z}),$$

then the CCA will produce biased results. A practical example of MNAR specified in this way may be a case-cohort study of homelessness and disease using frozen tissue samples. Suppose

self-reported homelessness status was recorded before tissue collection and homeless subjects are less likely to report their housing status. Further suppose an assay for a specific covariate was only performed at random on every other sample from subjects in the cohort that had an event and known housing status. Missingness of the covariate in the assay is MNAR and also depends on the outcome of interest.

## 4. Procedures for checking consistency with CIMAR and FIMAR assumptions

Complete case analysis (CCA) is the most used method for researchers handling missing covariate data in Cox models.[12] This is probably because most popular statistical programs (SAS, R, STATA) use CCA as the default method when performing Cox regression. Unfortunately, CCA relies on MCAR assumptions to produce reasonable parameter estimates. However, due to the non-identifiability issue one cannot verify the underlying missing mechanism.[2,11] Even if one cannot verify the underlying missing mechanism, under independent censoring and CIMAR/FIMAR working assumptions, Rathouz shows that one can use the observed data to check whether the missingness pattern is consistent with CIMAR/FIMAR using a two-step procedure.[4] This makes CIMAR and FIMAR appealing as researchers can check whether missingness of covariate data in survival analyses is consistent with their assumptions under independent censoring. No such procedure(s) exist for MCAR and MAR assumptions. I describe the two-step procedure below:

*Step 1: Treating C as failure time and T as censoring time in the subset for which X is observed, evaluate whether C is independent of X given Z.*

This step is based on the following property established by Rathouz. Assuming either CIMAR or FIMAR as a working assumption, CIMAR and $C \perp\!\!\!\perp X|Z$ together imply $C \perp\!\!\!\perp T|(R = 1, X, Z)$ and $C \perp\!\!\!\perp X|(R = 1, Z)$. Similarly, FIMAR and $C \perp\!\!\!\perp X|Z$ together imply $C \perp\!\!\!\perp T|(R = 1, X, Z)$ and $C \perp\!\!\!\perp X|(R = 1, Z)$. This indicates one can evaluate whether censoring time C is independent of the missing covariate X given Z based on the complete cases (i.e. R = 1).[4]

*Step 2: If so, evaluate whether C is independent of R given Z, which should hold under CIMAR, or whether T is independent of R given Z, which should hold under FIMAR.*

This step is based on the following property established in Rathouz. CIMAR and $C \perp\!\!\!\perp X|Z$ together imply $C \perp\!\!\!\perp T|R,Z$ and $C \perp\!\!\!\perp R|Z$. This indicates that under $C \perp\!\!\!\perp X|Z$ one can test CIMAR by modeling $C$ as a function of $(R, Z)$ by treating $T$ as the censoring time. Similarly, FIMAR and $C \perp\!\!\!\perp X|Z$ together imply $C \perp\!\!\!\perp T|R,Z$ and $T \perp\!\!\!\perp R|Z$. So, under $C \perp\!\!\!\perp X|Z$, one can test FIMAR by modeling $T$ as a function of $(R, Z)$, treating $C$ as the censoring time.[4] For both steps, Cox regression can evaluate the dependence.

## 5. Simulation

Since covariate missingness assumptions are unverifiable in real-world survival datasets, a simulation study is the best way to investigate how CIMAR and FIMAR may affect regression estimates in complete case Cox regression. With simulated data, I directly specify all parameters and I know precisely how covariates are missing. First, I generate covariates from normal distributions. I then generate failure times and censoring times that are independent of each other but depend on these normally distributed covariates. Finally, I model the probability of missing covariate X under MCAR, MAR, CIMAR, FIMAR and MNAR mechanisms. I perform these steps 500 times with two different sample sizes at low, medium, and high proportions of missing covariate data and present the average estimates for comparison. I describe the simulation procedures in detail below.

### *5.1 Samples, sample size and normally distributed covariates*

I generate covariates (X, Z) for 500 samples of 100 observations and 500 samples of 300 observations. X and Z are normally distributed as follows:

$$X \sim normal(10, 1.5^2),$$

$$Z \sim normal(70, 4^2).$$

*5.2 Failure times and censoring times*

I generate failure times (T) and censoring times (C) independently from exponential distributions. Both X and Z contribute to failure times and censoring times: X increases the hazard rate by 1.5 ($\beta_1 = \log(1.5)$) and Z increases the hazard rate by 1.10 ($\beta_2 = \log(1.10)$) in a per unit manner. T and C are distributed as follows:

$$\text{T} \sim 100000 * \exp(1 * 1.5^X * 1.10^Z) \approx \text{T} \sim 100000 * \exp(1 * e^{0.405*X+0.0953*Z}),$$

$$\text{C} \sim 100000 * \exp(0.25 * 1.5^X * 1.10^Z) \approx \text{T} \sim 100000 * \exp(0.25 * e^{0.405*X+0.0953*Z}).$$

Note: $\log(1.5) \approx 0.405$ and $\log(1.10) \approx 0.0953$. C is proportional to T such that approximately 20% of all observations should be censored (without missing data).

*5.3 Missing mechanisms and rates*

I assign missingness of covariate X (R) using Bernoulli distributions such that $\Pr(missing\ X)$ is MCAR, FIMAR, CIMAR, MAR or MNAR with missing rates of 15%, 45%, and 75%. I use the logit link function to specify $\Pr(missing\ X)$ as it connects covariates (X, Z) and survival data (Y, T, C, D) with covariate missingness (R), a property essential to assumptions of the various missingness mechanisms (see Appendix). Given covariates and survival data, I present the following general expression for $\Pr(missing\ X)$:

$$P = \frac{e^{\beta_0+\beta_1 Z+\beta_2 D+\beta_3 T+\beta_4 C+\beta_5 Y+\beta_6 X}}{1+e^{\beta_0+\beta_1 Z+\beta_2 D+\beta_3 T+\beta_4 C+\beta_5 Y+\beta_6 X}},$$

where P is $\Pr(missing\ X)$ and $\beta's$ take on some value depending on the covariate missingness mechanism and missing rate. If $\Pr(missing\ X)$ does not depend on some variable (covariates or survival data), then $\beta$ for that variable is 0 (e.g. $\beta_3$ for FIMAR simulation is 0 since the missingness

of covariates which are FIMAR does not depend on T). Using this general expression, I specify $\Pr(missing\ X)$ for the various covariate missingness mechanisms as follows:

Table A: $\Pr(missing\ X)$ under various covariate missingness mechanisms at 15%, 45%, and 75% missing.

| Missingness Mechanism | | MCAR | CIMAR |
|---|---|---|---|
| Definition | | $R \coprod (\mathrm{Y, D, X}) \mid \mathrm{Z}$ | $R \coprod (\mathrm{C, X}) \mid (\mathrm{T, Z})$ |
| Proportion of missing covariates | 15% | $\frac{e^{\log\left(\frac{1}{19}\right)+\log\left(\frac{57}{17}\right)*\left(\frac{1}{70}\right)*Z}}{1+e^{\log\left(\frac{1}{19}\right)+\log\left(\frac{57}{17}\right)*\left(\frac{1}{70}\right)*Z}}$ | $\frac{e^{\log\left(\frac{1}{19}\right)+0.327*t}}{1+e^{\log\left(\frac{1}{19}\right)+0.327*t}}$ |
| | 45% | $\frac{e^{\log\left(\frac{1}{19}\right)+\log\left(\frac{171}{11}\right)*\left(\frac{1}{70}\right)*Z}}{1+e^{\log\left(\frac{1}{19}\right)+\log\left(\frac{171}{11}\right)*\left(\frac{1}{70}\right)*Z}}$ | $\frac{e^{\log\left(\frac{1}{19}\right)+1.453*t}}{1+e^{\log\left(\frac{1}{19}\right)+1.453*t}}$ |
| | 75% | $\frac{e^{\log\left(\frac{1}{19}\right)+\log(57)*\left(\frac{1}{70}\right)*Z}}{1+e^{\log\left(\frac{1}{19}\right)+\log(57)*\left(\frac{1}{70}\right)*Z}}$ | $\frac{e^{\log\left(\frac{1}{19}\right)+5.23*t}}{1+e^{\log\left(\frac{1}{19}\right)+5.23*t}}$ |
| Missingness Mechanism | | FIMAR | MAR |
| Definition | | $R \coprod (\mathrm{T, X}) \mid (\mathrm{C, Z})$ | $R \coprod \mathrm{X} \mid (\mathrm{Y, D, Z})$ |
| Proportion of missing covariates | 15% | $\frac{e^{\log\left(\frac{1}{19}\right)+0.00821*c}}{1+e^{\log\left(\frac{1}{19}\right)+0.00821*c}}$ | $\frac{e^{\log\left(\frac{1}{19}\right)+0.005*Z+0.5*D+0.01625*time}}{1+e^{\log\left(\frac{1}{19}\right)+0.005*Z+0.5*D+0.01625*time}}$ |
| | 45% | $\frac{e^{\log\left(\frac{1}{19}\right)+0.03633*c}}{1+e^{\log\left(\frac{1}{19}\right)+0.03633*c}}$ | $\frac{e^{\log\left(\frac{1}{19}\right)+0.01*Z+1*D+0.0684*time}}{1+e^{\log\left(\frac{1}{19}\right)+0.01*Z+1*D+0.0684*time}}$ |
| | 75% | $\frac{e^{\log\left(\frac{1}{19}\right)+0.1309*c}}{1+e^{\log\left(\frac{1}{19}\right)+0.1309*c}}$ | $\frac{e^{\log\left(\frac{1}{19}\right)+0.015*Z+1.5*D+0.1868*time}}{1+e^{\log\left(\frac{1}{19}\right)+0.015*Z+1.5*D+0.1868*time}}$ |
| Missingness Mechanism | | MNAR1 | MNAR2 |
| Definition | | $R \mid (\mathrm{X, Z})$ | $R \mid (\mathrm{X, Y, D, Z})$ |
| Proportion of missing covariates | 15% | $\frac{e^{\log\left(\frac{1}{19}\right)+0.1195*X}}{1+e^{\log\left(\frac{1}{19}\right)+0.1195*X}}$ | $\frac{e^{\log\left(\frac{1}{19}\right)+0.005*Z+0.5*D+0.008125*time+0.0236*X}}{1+e^{\log\left(\frac{1}{19}\right)+0.005*Z+0.5*D+0.08125*time+0.0236*X}}$ |
| | 45% | $\frac{e^{\log\left(\frac{1}{19}\right)+0.2738*X}}{1+e^{\log\left(\frac{1}{19}\right)+0.2738*X}}$ | $\frac{e^{\log\left(\frac{1}{19}\right)+0.01*Z+1*D+0.0342*time+0.054*X}}{1+e^{\log\left(\frac{1}{19}\right)+0.01*Z+1*D+0.0342*time+0.054*X}}$ |
| | 75% | $\frac{e^{\log\left(\frac{1}{19}\right)+0.4134*X}}{1+e^{\log\left(\frac{1}{19}\right)+0.4134*X}}$ | $\frac{e^{\log\left(\frac{1}{19}\right)+0.015*Z+1.5*D+0.0934*time+0.066*X}}{1+e^{\log\left(\frac{1}{19}\right)+0.015*Z+1.5*D+0.0934*time+0.066*X}}$ |

Where $X$ is the missing covariate, $R = I(X\ is\ observed)$, $T = failure\ time$, $C = censoring\ time$, $time = Y = \min(T, C)$, $D = I(T \leq C)$, and $Z$ is some other covariate without missing data; $I(\cdot)$ is an indicator function.

I use an intercept ($\beta_0$) of $\log\left(\frac{1}{19}\right)$ such that at least 5% of all missing data is missing completely at random. There is no closed form for these expressions and the parameter values for the covariates were determined in R with discretion such that all variables relevant to the covariate missingness mechanism contribute significantly to the missing rate (e.g. Y and D both have significant effects on the covariate missing rate in the simulated MAR data). I perform all simulation in R (R Studio) with the aid of the statistical packages 'OIsurv', 'broom' and 'boot'.

*5.4 Simulation results*

Table 1: Complete case analysis results. Low proportion of missing covariates (15%). Average of 500 samples with 100 subjects.

| Data (Missing) | $\beta_1$ | | | | | $\beta_2$ | | | | | Censor |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Estimate | Bias (%) | SD | SE | Coverage 95% | Estimate | Bias (%) | SD | SE | Coverage 95% | |
| Fully Observed | 0.4177 | 0.0122 (3.01) | 0.0880 | 0.0894 | 0.958 | 0.0968 | 0.0015 (1.61) | 0.0311 | 0.0311 | 0.958 | 0.201 |
| MCAR (0.150) | 0.4197 | 0.0142 (3.51) | 0.0971 | 0.0980 | 0.958 | 0.0965 | 0.0012 (1.28) | 0.0343 | 0.0342 | 0.952 | 0.201 |
| CIMAR (0.147) | 0.3525 | -0.0529 (-13.05) | 0.0953 | 0.0977 | 0.926 | 0.0817 | -0.0136 (-14.30) | 0.0344 | 0.0337 | 0.932 | 0.178 |
| FIMAR (0.147) | 0.4212 | 0.0158 (3.88) | 0.1025 | 0.1009 | 0.956 | 0.0971 | 0.0018 (1.84) | 0.0336 | 0.0348 | 0.97 | 0.221 |
| MAR (0.148) | 0.3893 | -0.0162 (-4.00) | 0.0961 | 0.0989 | 0.94 | 0.0902 | -0.0051 (-5.34) | 0.0344 | 0.0343 | 0.948 | 0.211 |
| MNAR1 (0.149) | 0.4196 | 0.0142 (3.49) | 0.0963 | 0.0980 | 0.958 | 0.0964 | 0.0011 (1.18) | 0.0345 | 0.0341 | 0.95 | 0.202 |
| MNAR2 (0.148) | 0.4057 | 0.0002 (0.05) | 0.0975 | 0.0986 | 0.956 | 0.0937 | -0.0017 (-1.74) | 0.0346 | 0.0344 | 0.948 | 0.212 |

Table 2: Complete case analysis results. Medium proportion of missing covariates (45%). Average of 500 samples with 100 subjects.

| Data (Missing) | $\beta_1$ | | | | | $\beta_2$ | | | | | Censor |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Estimate | Bias (%) | SD | SE | Coverage 95% | Estimate | Bias (%) | SD | SE | Coverage 95% | |
| Fully Observed | 0.4177 | 0.0122 (3.01) | 0.0880 | 0.0894 | 0.958 | 0.0968 | 0.0015 (1.61) | 0.0311 | 0.0311 | 0.958 | 0.201 |
| MCAR (0.452) | 0.4241 | 0.0186 (4.59) | 0.1273 | 0.1263 | 0.956 | 0.0987 | 0.0034 (3.55) | 0.0477 | 0.0444 | 0.936 | 0.203 |
| CIMAR (0.448) | 0.1946 | -0.2109 (-52.01) | 0.1214 | 0.1181 | 0.522 | 0.0489 | -0.0464 (-48.73) | 0.0440 | 0.0411 | 0.77 | 0.118 |
| FIMAR (0.447) | 0.4259 | 0.0204 (5.03) | 0.1460 | 0.1392 | 0.946 | 0.1004 | 0.0051 (5.36) | 0.0490 | 0.0483 | 0.952 | 0.316 |
| MAR (0.450) | 0.2755 | -0.1300 (-32.06) | 0.1353 | 0.1287 | 0.786 | 0.0652 | -0.0301 (-31.62) | 0.0458 | 0.0445 | 0.894 | 0.249 |
| MNAR1 (0.448) | 0.4246 | 0.0191 (4.72) | 0.1342 | 0.1274 | 0.948 | 0.0982 | 0.0029 (3.08) | 0.0451 | 0.0438 | 0.95 | 0.202 |
| MNAR2 (0.450) | 0.3269 | -0.0786 (-19.38) | 0.1305 | 0.1305 | 0.9 | 0.0763 | -0.0190 (-19.99) | 0.0460 | 0.0451 | 0.93 | 0.260 |

Table 3: Complete case analysis results. High proportion of missing covariates (75%). Average of 500 samples with 100 subjects.

| Data (Missing) | $\beta_1$ | | | | | $\beta_2$ | | | | | Censor |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Estimate | Bias (%) | SD | SE | Coverage 95% | Estimate | Bias (%) | SD | SE | Coverage 95% | |
| Fully Observed | 0.4177 | 0.0122 (3.01) | 0.0880 | 0.0894 | 0.958 | 0.0968 | 0.0015 (1.61) | 0.0311 | 0.0311 | 0.958 | 0.201 |
| MCAR (0.747) | 0.4487 | 0.0433 (10.67) | 0.2472 | 0.2091 | 0.958 | 0.1021 | 0.0068 (7.17) | 0.0819 | 0.0729 | 0.934 | 0.202 |
| CIMAR (0.749) | 0.0885 | -0.3170 (-78.17) | 0.2002 | 0.1771 | 0.522 | 0.0215 | -0.0738 (-77.44) | 0.0678 | 0.0648 | 0.768 | 0.055 |
| FIMAR (0.751) | 0.4812 | 0.0758 (18.68) | 0.3581 | 0.2764 | 0.948 | 0.1109 | 0.0156 (16.35) | 0.1216 | 0.0979 | 0.94 | 0.503 |
| MAR (0.750) | 0.1730 | -0.2325 (-57.33) | 0.2561 | 0.2192 | 0.74 | 0.0363 | -0.0590 (-61.87) | 0.0855 | 0.0787 | 0.858 | 0.334 |
| MNAR1 (0.752) | 0.4580 | 0.0526 (12.97) | 0.2469 | 0.2198 | 0.932 | 0.1050 | 0.0097 (10.14) | 0.0808 | 0.0731 | 0.938 | 0.203 |
| MNAR2 (0.751) | 0.2186 | -0.1869 (-46.10) | 0.2575 | 0.2271 | 0.818 | 0.0522 | -0.0431 (-45.25) | 0.0921 | 0.0800 | 0.908 | 0.361 |

Data generated under CIMAR assumptions exhibit significant bias in CCA, even when the proportion of missing data is small. At 15% missing under CIMAR assumptions, estimates for both $\beta_1$ and $\beta_2$ exhibit significant bias compared to fully observed data and MCAR data. Estimates for $\beta_1$ and $\beta_2$ have biases of -13.05% and -14.30% respectively for CIMAR data compared to biases of 3.01% and 1.61% for fully observed data and biases of 3.51% and 1.28% for MCAR data. In addition, the coverage rates, the proportion of samples that contain the true population parameter in the 95% confidence interval for the estimates, are 0.926 and 0.932 respectively (Table 1). At 45% missing, estimates for $\beta_1$ and $\beta_2$ have biases of -52.01% and -48.73% for CIMAR data compared to biases of 4.59% and 3.55% for MCAR data. The coverages rates are 0.522 and 0.77 (Table 2). At 75% missing, estimates for $\beta_1$ and $\beta_2$ have biases of -78.17% and -77.44% for CIMAR data compared to biases of 10.67% and 7.17% for MCAR data. The coverages rates are 0.522 and 0.768 (Table 3). Censoring rates for CIMAR data are lower than MCAR data (Table 1-3). This was expected as I specify that longer failure times increase $\Pr(missing\ X)$ which selects for the presence of subjects with shorter failure times. As subjects with shorter failure times are less likely to be censored at random, the censoring rate decreased. These results are consistent with survival

literature suggesting an increasing censoring rate increases variance and uneven censoring patterns can lead to biased estimates.[13]

Note that while survival datasets with missing covariates which are MCAR are known to produce consistent parameter estimates in complete case Cox regression, the estimates are not completely unbiased. The Cox model produces asymptotically unbiased estimates in CCA with MCAR data under independent censoring.[7] With finite datasets such as this simulation, it is not unusual to observe some small bias in the results of complete case Cox regression with missing covariates that are MCAR, especially with high missing rates.

Data generated under MAR assumptions exhibit significant bias with medium and large proportions of missing data. At 15% missing under MAR assumptions, estimates for both $\beta_1$ and $\beta_2$ did not exhibit significant bias compared to fully observed data and MCAR data. Estimates for $\beta_1$ and $\beta_2$ have biases of -4.00% and -5.34% respectively for MAR data, comparable to biases of 3.01% and 1.61% for fully observed data and biases of 3.51% and 1.28% for MCAR data. The coverage rates are 0.94 and 0.948 respectively (Table 1). At 45% missing, estimates for $\beta_1$ and $\beta_2$ have biases of -32.06% and -31.62% for MAR data compared to biases of 4.59% and 3.55% for MCAR data. The coverages rates are 0.786 and 0.894 (Table 2). At 75% missing, estimates for $\beta_1$ and $\beta_2$ have biases of -57.33% and -61.87% for MAR data compared to biases of 10.67% and 7.17% for MCAR data. The coverages rates are 0.74 and 0.858 (Table 3). Censoring rates for MAR data are higher than MCAR data (Table 1-3). This was expected as I specify failure status $(D = 1)$ increases $\Pr(missing\ X)$ which selects for subjects with censored survival time. As subjects who fail are more likely to be missing, the censored rate increases.

Data generated using FIMAR assumptions did not exhibit significant bias, even with a large proportion of missing data. At 15% missing, estimates for both $\beta_1$ and $\beta_2$ have biases of 3.88% and 1.84% respectively for FIMAR data, similar to biases of 3.01% and 1.61% for fully observed data and biases of 3.51% and 1.28% for MCAR data. The coverages rates are 0.956 and 0.97 respectively (Table 1). At 45% missing, estimates for $\beta_1$ and $\beta_2$ have biases of 5.03% and 5.36% for FIMAR compared to biases of 4.59% and 3.55% for MCAR data. The coverages rates are 0.946 and 0.952 (Table 2). At 75% missing, estimates for $\beta_1$ and $\beta_2$ have biases of 18.68% and 16.35% for FIMAR data compared to biases of 10.67% and 7.17% for MCAR data. The coverages rates are 0.948 and 0.94 (Table 3). Censoring rates for FIMAR data are higher than MCAR data (Table 1-3). This was expected as I specify that longer censoring times increase $\Pr(missing\ X)$ which selects for the presence of subjects with shorter censoring times. As subjects with shorter censoring times are more likely to be censored at random, the censoring rate increases.

Data generated using MNAR assumptions dependent on survival data (MNAR2) exhibit significant bias at 45% and 75% missing whereas data generated using MNAR assumptions independent of survival data (MNAR1) did not. At 15% missing data under MNAR2 assumptions, estimates for both $\beta_1$ and $\beta_2$ did not exhibit significant bias compared to fully observed data and MCAR data. Estimates for $\beta_1$ and $\beta_2$ have biases of 0.05% and -1.74% respectively for MNAR2 data compared to biases of 3.01% and 1.61% for fully observed data and biases of 3.51% and 1.28% for MCAR data. The coverage rates are 0.956 and 0.948 respectively (Table 1). At 45% missing, estimates for $\beta_1$ and $\beta_2$ have biases of -19.38% and -19.99% for MNAR2 data compared to biases of 4.59% and 3.55% for MCAR data. The coverages rates are 0.9 and 0.93 (Table 2). At 75% missing, estimates for $\beta_1$ and $\beta_2$ have biases of -46.10% and -45.25% for MNAR2 data compared to biases of 10.67% and 7.17% for MCAR data. The coverages rates are 0.818 and 0.908

(Table 3). Censoring rates for MNAR2 data are higher than MCAR data (Table 1-3). This was expected as I specify failure status $(D = 1)$ increases $\Pr(missing\ X)$ which selects for subjects with censored survival time. As subjects who fail are more likely to be missing, the censored rate increases.

Data generated using MNAR assumptions independent of survival data (MNAR1) did not exhibit significant bias, even with a large proportion of missing data. At 15% missing, estimates for both $\beta_1$ and $\beta_2$ have biases of 3.49% and 1.18% respectively for MNAR1 data, similar to biases of 3.01% and 1.61% for fully observed data and biases of 3.51% and 1.28% for MCAR data. The coverages rates are 0.958 and 0.95 respectively (Table 1). At 45% missing, estimates for $\beta_1$ and $\beta_2$ have biases of 4.72% and 3.08% for MNAR1 data, similar to biases of 4.59% and 3.55% for MCAR data. The coverages rates are 0.948 and 0.95 (Table 2). At 75% missing, estimates for $\beta_1$ and $\beta_2$ have biases of 12.97% and 10.14% for MNAR1 data, similar to biases of 10.67% and 7.17% for MCAR data. The coverages rates are 0.932 and 0.938 (Table 3).

Simulation results of MNAR dependent on survival data (MNAR2) and MNAR independent of survival data (MNAR1) are consistent with literature. As mentioned previously, survival data where missingness depends only on covariates (both present and missing) but not survival data yields reasonable estimators in complete case Cox regression.[4] As such, we expect CCA of MNAR1 data to produce reasonable parameter estimates and CCA of MNAR2 data to produce biased parameter estimates in Cox regression as is the case in this simulation study.

In addition to determining whether survival datasets with missing covariates under various mechanisms are appropriate for complete case Cox regression, I assess which missingness mechanisms produce the "best" estimates. In parameter estimation, the "best" estimates are those

with the least (absolute) bias and least variance but there is a bias/variance tradeoff where bias and variance are inversely related.[14] Complete case Cox regression with missing covariate data generated under MCAR mechanisms, MNAR mechanisms independent of survival data (MNAR1), and FIMAR mechanisms all produced parameter estimates with negligible or slight bias. FIMAR results are similar to MCAR results and MNAR1 results, both of which exhibit reasonable parameter estimates. This differs from CIMAR results, which exhibit significantly more bias than results from MCAR or FIMAR data; CIMAR results exhibit the most bias of any missingness mechanism explored in this paper (Table 4).

Table 4: $\beta_1$ Estimates ordered by bias. 100 subjects.

| Missing % | Bias | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 Least bias | 2 | 3 | 4 | 5 | 6 | 7 Most bias |
| 15% | MNAR2 | Fully observed | MNAR1 | MCAR | FIMAR | MAR | CIMAR |
| 45% | Fully observed | MCAR | MNAR1 | FIMAR | MNAR2 | MAR | CIMAR |
| 75% | Fully observed | MCAR | MNAR1 | FIMAR | MNAR2 | MAR | CIMAR |

In contrast to bias, FIMAR results exhibit the most variance compared to the other missingness mechanisms at 45% and 75% missing (Table 5). This is not surprising given the bias-variance tradeoff described above. CIMAR results exhibit the least variance (besides fully observed data) and the most bias. There are negligible differences in variance between results of the various missing mechanisms at 15% missing (Table 1 and Table 5). While FIMAR did have the most variance and CIMAR the least, the differences between any of the variances from results for the different covariate missing mechanisms were not large (e.g. with 75% missing, the SD for $\beta_1$ are 0.2002, 0.2472 and 0.3581 respectively for CIMAR, MCAR, and FIMAR data; Table 3).

FIMAR results exhibit increased variance compared to MCAR results but not enough to suggest FIMAR data are inappropriate for complete case Cox regression.

Table 5: $\beta_1$ Estimates ordered by variance. 100 subjects.

| Missing % | Variance | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 Least variance | 2 | 3 | 4 | 5 | 6 | 7 Most variance |
| 15% | Fully observed | CIMAR | MCAR* | MNAR1* | MNAR2 | MAR | FIMAR |
| 45% | Fully observed | CIMAR | MCAR | MNAR1 | MAR | MNAR2 | FIMAR |
| 75% | Fully observed | CIMAR | MCAR | MAR | MNAR1 | MNAR2 | FIMAR |

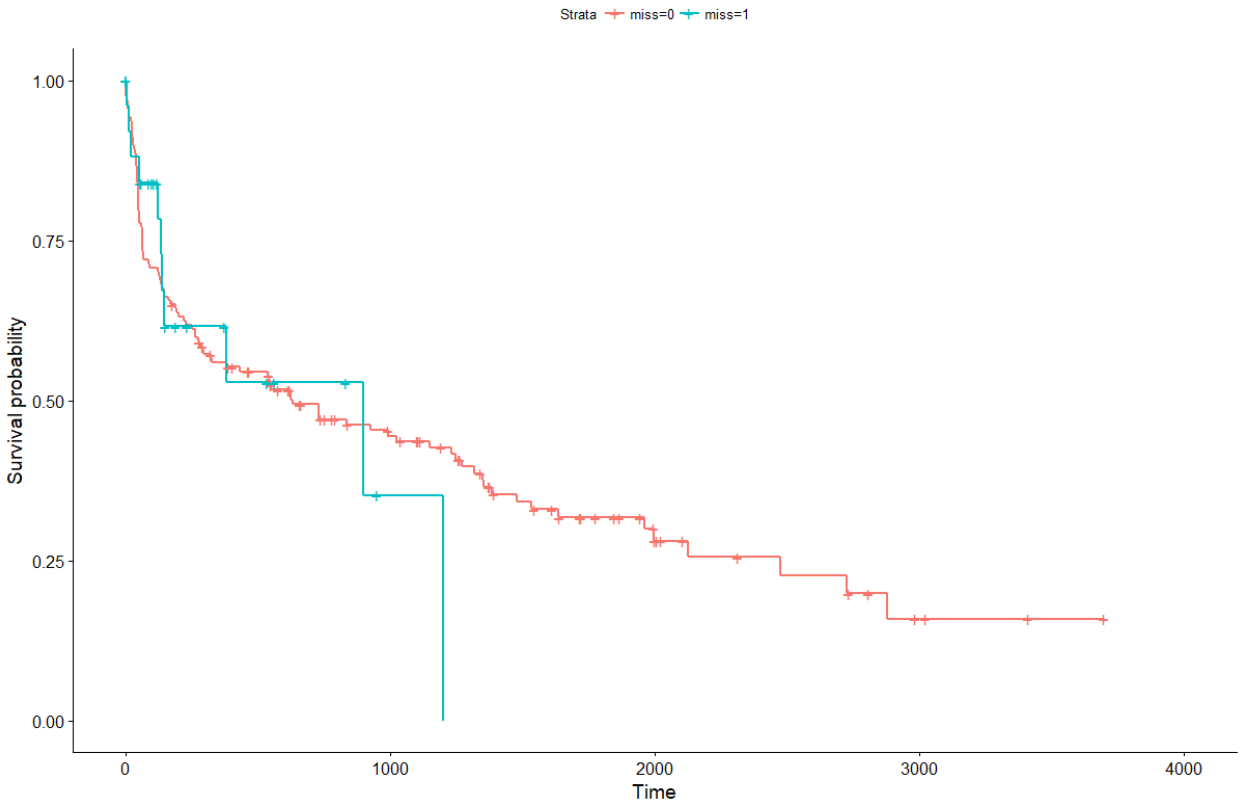
*these two variances are approximately the same for this simulation with 15% missing covariates.

These findings are consistent with results from the simulated samples with 300 subjects each (see Tables A1, A2, and A3 in the Appendix).

## 6. Assessing CIMAR and FIMAR in real data

I apply Rathouz's two-step procedure to the Stanford heart transplant survival dataset (https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/csv/survival/stanford2.csv). Researchers recorded the survival/censoring time, censoring status, age, and T5 mismatch score, a measure of dissimilarity between HL-A antigens of donors and recipients. Some patients have missing T5 mismatch scores; all other data were fully observed.[15] I treat T5 mismatch score as the missing covariate of interest (X) with age as some other covariate (Z). I perform all analyses in R (R Studio) with the aid of the statistical packages 'OIsurv', 'survival' and 'survminer'.

Figure 1: Estimated survival curve for heart transplant patients by T5 mismatch score missingness status




Survival for patients with known T5 scores in red. Survival for patients with unknown T5 scores in blue. Time in days. Crosses indicate censored events.

Table 6: Results of Steps 2 and 3 of Rathouz's procedure for Stanford heart transplant data.

| N | $N_{missing}$ % | Step 1 | | Step 2 | | | |
|---|---|---|---|---|---|---|---|
| | | $HR_1$ | $p_1$ | $HR_2$ | $p_2$ | $HR_3$ | $p_3$ |
| 184 | 27 (14.67) | 1.3096 | 0.3300 | 0.1183 | 0.0000 | 0.9185 | 0.7919 |

$HR_2$ and $p_2$ correspond with Step 2 for CIMAR. $HR_3$ and $p_3$ correspond with Step 2 for FIMAR.

157 patients of whom 102 (65.0%) fail have known T5 mismatch scores while 27 patients of whom 11 (40.7%) fail have missing T5 mismatch scores. Survival is similar in patients with known T5 scores and patients with missing T5 scores until ~1000 days. Censoring events occur earlier in the patients with missing T5 scores (Figure 1). Given the similar survival curves for subjects with known T5 mismatch scores and subjects with unknown T5 mismatch scores as well as the observation that censoring events occur earlier in patients with unknown T5 mismatch scores, I suggest FIMAR as a working assumption.

Treating C as failure time and T as censoring time in the subset for which T5 mismatch scores are observed, C is independent of X ($p = 0.3300$) given Z. Step 1 is satisfied, missingness is consistent with CIMAR or FIMAR. Treating T as censoring time and modeling C as a function of (R, Z), C is dependent on R ($p < 0.0001$) (Table 6). Missingness is not consistent with CIMAR. Treating C as censoring time and modeling T as a function of (R, Z), T is independent of R ($p = 0.7919$) (Table 6). Missingness is consistent with FIMAR. Therefore FIMAR is a reasonable potential missingness mechanism for the Stanford heart transplant datasets and CCA should yield reasonable parameter estimates. Complete case Cox regression of the Stanford heart transplant dataset suggests that T5 mismatch scores did not have a significant effect on patient survival times ($p = 0.3545$); age had a small (HR = 1.03) but significant ($p = 0.0092$) effect on survival.

## 7. Discussion

Simulation results suggest that survival data with missing covariates that are influenced by failure times (T) are not appropriate for complete case Cox regression. In simulation, CIMAR data and MAR data dependent on survival data (Y, D) produce unacceptably biased parameter estimates compared to MCAR data in CCA. Conversely, FIMAR data produces reasonable though possibly inefficient parameter estimates in complete case Cox regression similar to MCAR data. In addition, CCA of MNAR data that depends on survival data (MNAR2) produces biased parameter estimates while CCA of MNAR data that does not depend on survival data (MNAR1) produces reasonable parameter estimates.

For appropriate analysis of survival data with missing covariates via complete case Cox regression, it is important for researchers to properly identify the underlying missingness mechanism. Following Rathouz's procedure, it should be possible to identify the survival dataset as either CIMAR or FIMAR providing the dataset "passes" step 1 (C is independent of the missing covariate X given Z based on the complete cases (i.e. R = 1)); if it does not "pass" step 1, a sensitivity analysis may be appropriate. If the data is FIMAR (T is independent of R given Z), complete case Cox regression should be appropriate. If the data is CIMAR (C is independent of R given Z), complete case analysis might not be appropriate and a sensitivity analysis should be performed for comparison with the complete case results. If results of the primary complete case analysis differ significantly from results of the sensitivity analysis, the adjusted findings from the sensitivity analysis may be more appropriate to report.

One limitation of this simulation study is the use of exponential survival times. Survival times which do not follow exponential distributions might produce different findings. Fortunately, it has been shown that other distributions such as Weibull and Gompertz also generate appropriate

survival times and produce consistent parameter estimates in complete case Cox regression with MCAR data.[16] Another limitation is parameterization of missingness coefficients. Since there is no closed form to calculate the missing probability, I use software (R) and personal discretion in specifying missingness coefficients for $\Pr(missing\ X)$. While personal discretion is not uncommon for parameterization in simulation studies, this approach is not systematic. Results will vary depending on how researchers specify missingness parameters.

Additional limitations are independent censoring assumptions and inability to use Rathouz's procedure when survival data does not "pass" step 1. Independent censoring is a necessary assumption for Rathouz's procedure and for Cox models to produce reasonable parameter estimates in CCA with MCAR data. Unfortunately, independent censoring is often difficult or impossible to verify. Given independent censoring, the usefulness of Rathouz's procedure is limited to datasets which "pass" step 1. When the dataset does not "pass" step 1, Rathouz's procedure cannot be used to check consistency of missing covariates with CIMAR and FIMAR.

## Appendix

The logit function takes on the general form:[17]

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i,$$

where P is some probability; in this case Pr($missing\ X$). i is the subject index. $X_i$ can be covariates and/or survival data. With some algebra, P may be presented as a function of the covariates and survival data, $X_i$:

$$P = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_i X_i}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_i X_i}}.$$

Now I can specify Pr($missing\ X$) as a function of our covariates (X, Z) and/or survival data (Y, T, C, D) as necessary under the various missingness mechanisms.

Table A1: Complete case analysis results. Low proportion of missing covariates (15%). Average of 500 samples with 300 subjects.

| Data (Missing) | $\beta_1$ | | | | | $\beta_2$ | | | | | Censor |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Estimate | Bias (%) | SD | SE | Coverage 95% | Estimate | Bias (%) | SD | SE | Coverage 95% | |
| Fully Observed | 0.4071 | 0.0016 (0.41) | 0.0515 | 0.0499 | 0.954 | 0.0962 | 0.0009 (0.98) | 0.0166 | 0.0174 | 0.96 | 0.199 |
| MCAR (0.151) | 0.4069 | 0.0014 (0.35) | 0.0569 | 0.0543 | 0.946 | 0.0961 | 0.0008 (0.86) | 0.0180 | 0.0189 | 0.962 | 0.199 |
| CIMAR (0.149) | 0.3392 | -0.0662 (-16.33) | 0.0547 | 0.0544 | 0.75 | 0.0803 | -0.0150 (-15.78) | 0.0174 | 0.0188 | 0.89 | 0.177 |
| FIMAR (0.152) | 0.4066 | 0.0012 (0.29) | 0.0549 | 0.0561 | 0.95 | 0.0961 | 0.0008 (0.85) | 0.0188 | 0.0194 | 0.956 | 0.220 |
| MAR (0.150) | 0.3763 | -0.0292 (-7.19) | 0.0592 | 0.0550 | 0.886 | 0.0895 | -0.0059 (-6.15) | 0.0180 | 0.0191 | 0.95 | 0.209 |
| MNAR1 (0.150) | 0.4063 | 0.0009 (0.22) | 0.0562 | 0.0544 | 0.958 | 0.0959 | 0.0005 (0.57) | 0.0178 | 0.0189 | 0.962 | 0.199 |
| MNAR2 (0.150) | 0.3932 | -0.0123 (-3.02) | 0.0577 | 0.0548 | 0.928 | 0.0928 | -0.0025 (-2.59) | 0.0181 | 0.0190 | 0.968 | 0.209 |

Table A2: Complete case analysis results. Medium proportion of missing covariates (45%). Average of 500 samples with 300 subjects.

| Data (Missing) | $\beta_1$ | | | | | $\beta_2$ | | | | | Censor |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Estimate | Bias (%) | SD | SE | Coverage 95% | Estimate | Bias (%) | SD | SE | Coverage 95% | |
| Fully Observed | 0.4071 | 0.0016 (0.41) | 0.0515 | 0.0499 | 0.954 | 0.0962 | 0.0009 (0.98) | 0.0166 | 0.0174 | 0.96 | 0.199 |
| MCAR (0.451) | 0.4053 | -0.0001 (-0.03) | 0.0668 | 0.0682 | 0.96 | 0.0958 | 0.0005 (0.51) | 0.0226 | 0.0238 | 0.962 | 0.198 |
| CIMAR (0.450) | 0.1892 | -0.2163 (-53.35) | 0.0642 | 0.0645 | 0.088 | 0.0434 | -0.0519 (-54.48) | 0.0231 | 0.0226 | 0.364 | 0.119 |
| FIMAR (0.449) | 0.4128 | 0.0073 (1.81) | 0.0778 | 0.0757 | 0.954 | 0.0963 | 0.0010 (1.07) | 0.0272 | 0.0260 | 0.942 | 0.315 |
| MAR (0.451) | 0.2580 | -0.1475 (-36.38) | 0.0716 | 0.0705 | 0.414 | 0.0618 | -0.0335 (-35.15) | 0.0237 | 0.0245 | 0.734 | 0.247 |
| MNAR1 (0.452) | 0.4060 | 0.0006 (0.14) | 0.0737 | 0.0694 | 0.93 | 0.0957 | 0.0004 (0.42) | 0.0242 | 0.0238 | 0.94 | 0.198 |
| MNAR2 (0.449) | 0.3061 | -0.0993 (-24.50) | 0.0710 | 0.0706 | 0.682 | 0.0730 | -0.0223 (-23.42) | 0.0239 | 0.0248 | 0.844 | 0.260 |

Table A3: Complete case analysis results. High proportion of missing covariates (75%). Average of 500 samples with 300 subjects.

| Data (Missing) | $\beta_1$ | | | | | $\beta_2$ | | | | | Censor |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Estimate | Bias (%) | SD | SE | Coverage 95% | Estimate | Bias (%) | SD | SE | Coverage 95% | |
| Fully Observed | 0.4071 | 0.0016 (0.41) | 0.0515 | 0.0499 | 0.954 | 0.0962 | 0.0009 (0.98) | 0.0166 | 0.0174 | 0.96 | 0.199 |
| MCAR (0.747) | 0.4216 | 0.0161 (3.98) | 0.1056 | 0.1048 | 0.96 | 0.1007 | 0.0054 (5.67) | 0.0393 | 0.0367 | 0.936 | 0.199 |
| CIMAR (0.750) | 0.0854 | -0.3200 (-78.93) | 0.0924 | 0.0909 | 0.076 | 0.0176 | -0.0777 (-81.55) | 0.0366 | 0.0331 | 0.336 | 0.058 |
| FIMAR (0.751) | 0.4194 | 0.0140 (3.45) | 0.1381 | 0.1343 | 0.94 | 0.0972 | 0.0018 (1.94) | 0.0490 | 0.0466 | 0.948 | 0.505 |
| MAR (0.750) | 0.1536 | -0.2519 (-62.12) | 0.1111 | 0.1104 | 0.366 | 0.0358 | -0.0595 (-62.41) | 0.0417 | 0.0394 | 0.65 | 0.333 |
| MNAR1 (0.749) | 0.4233 | 0.0178 (4.40) | 0.1125 | 0.1084 | 0.944 | 0.0995 | 0.0042 (4.38) | 0.0401 | 0.0366 | 0.924 | 0.199 |
| MNAR2 (0.750) | 0.2191 | -0.1863 (-45.95) | 0.1171 | 0.1142 | 0.584 | 0.0453 | -0.0500 (-52.42) | 0.0406 | 0.0399 | 0.76 | 0.363 |

Please contact me at minzhu@email.arizona.edu with regards to simulation code, questions, or comments.

## References


1. George, B., Seals, S., & Aban, I. (2014). Survival analysis and regression models. *Journal of Nuclear Cardiology*, *21*(4), 686-694.

2. Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 222.

3. Ali, A. M. G., Dawson, S. J., Blows, F. M., Provenzano, E., Ellis, I. O., Baglietto, L., ... & Pharoah, P. D. (2011). Comparison of methods for handling missing data on immunohistochemical markers in survival analysis of breast cancer. *British journal of cancer*, *104*(4), 693.

4. Rathouz, P. J. (2006). Identifiability assumptions for missing covariate data in failure time regression models. *Biostatistics*, *8*(2), 345-356.

5. Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data* (Vol. 333). John Wiley & Sons.

6. Paik, M. C., & Tsai, W. Y. (1997). On using the Cox proportional hazards model with missing covariates. *Biometrika*, *84*(3), 579-593.

7. Cox, D. R. (1975). Partial likelihood. *Biometrika*, *62*(2), 269-276.

8. Cox, D. R. (2018). *Analysis of survival data*. Routledge.

9. Fox, J. (2002). Cox proportional-hazards regression for survival data. *An R and S-PLUS companion to applied regression*, *2002*.

10. Marshall, A., Altman, D. G., Royston, P., & 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), 7.

11. Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

12. Burton, A., & Altman, D. G. (2004). Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *British Journal of Cancer*, *91*(1), 4.

13. Buckley, J., & James, I. (1979). Linear regression with censored data. *Biometrika*, *66*(3), 429-436.

14. Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, *4*(1), 1-58.

15. Andrews, D. F., & Herzberg, A. M. (1985). Stanford Heart Transplant Data. In *Data* (pp. 45-50). Springer, New York, NY.

16. Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine*, *24*(11), 1713-1723.

17. Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Pacific Grove, CA: Duxbury.