

FRAILTY MODEL INCORPORATING ASCERTAINMENT CORRECTION WITH MISSING  
DATA IN FAMILY-BASED STUDY

by

Jiaqi Bi

A thesis submitted in conformity with the requirements  
for the degree of Master of Science  
Graduate Department of Epidemiology and Biostatistics  
Schulich School of Medicine & Dentistry  
University of Western Ontario

© Copyright 2024 by Jiaqi Bi

# **Abstract**

Frailty Model Incorporating Ascertainment Correction with Missing Data in Family-Based  
Study

Jiaqi Bi

Master of Science

Graduate Department of Epidemiology and Biostatistics

Schulich School of Medicine & Dentistry

University of Western Ontario

2024

This is an abstract section...

*To my mother, Kai Hua, for all the upbringings.*  
*To my father, Guangmin Bi, for teaching me not to give up.*  
*To my supervisors, for their unstopping guidance and invaluable lessons.*

## Acknowledgements

# Contents

<b>Abstract</b>	<b>ii</b>
<b>Dedication</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>List of Symbols</b>	<b>vi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Motivation . . . . .	2
1.3 Objectives . . . . .	2
1.4 Organizations of the Thesis . . . . .	3
<b>2 Literature Reviews</b>	<b>4</b>
2.1 Survival Analysis . . . . .	4
2.1.1 Frailty Model for Family Based Study . . . . .	5
<b>3 Methods</b>	<b>6</b>
3.1 Introduction . . . . .	6
3.2 Kinship Matrix . . . . .	6
<b>Bibliography</b>	<b>7</b>

# List of Symbols

$i$	Individual index
$j$	Family (Cluster) index
$p$	Proband index
$d_j$	Number of events in family $j$
$t$	Some time
$a$	Some Time for the proband
$T$	Event Time
$\delta_{ij}$	Event indicator for individual $i$ in family $j$
$w$	The observed survival data $(t, \delta)$
$n$	Number of individuals
$J$	Number of Families (Clusters)
$m$	Index of the sampled completed dataset in the MCEM
$M$	Number of the sampled completed dataset in the MCEM
$z$	Frailty term
$q$	$q$ -th element of Gauss Hermite Quadrature
$\omega$	$q$ -th weight of Gauss Hermite Quadrature
$y_q$	$q$ -th node of Gauss Hermite Quadrature
$N_q$	Total number of quadratures
$h(\cdot)$	Hazard fucntion
$h_0(\cdot)$	Baseline hazard function
$H(\cdot)$	Cumulative hazard fucntion
$S(\cdot)$	Survival fucntion
$A_j(\cdot)$	Ascertainment of family $j$ into the study
$L(\cdot)$	Likelihood function
$\ell(\cdot)$	Log-likelihood function
$\mathcal{L}(\cdot)$	Laplace transform
$\mathbf{x}$	Covariates
$\boldsymbol{\beta}$	Model coefficients vector
$\boldsymbol{\theta}$	Parameter vector
$\Lambda$	The combination of $(\boldsymbol{\beta}, \lambda, \alpha)$

$\lambda$	Weibull shape parameter
$\alpha$	Weibull scale parameter
$v$	General form of the parameter in an undefined frailty distribution
$k$	Gamma shape and rate parameters
$\sigma^2$	Log-Normal variance parameter
$\psi$	Missing data distribution parameters

# List of Tables



# List of Figures

# Chapter 1

## Introduction

### 1.1 Background

The Breast Cancer type 1/2, usually referred as BRCA1/2, are proteins that consists of genes that code for BRCA1 in humans. BRCA1/2 are human tumor suppressor genes, that are responsible for repairing the DNA [1]. When the mutation exists on these genes may cause the impairments of proper functions, which can lead to the possibility of capturing the breast, ovarian, or other specific cancers [2, 3]. Inheriting one of these mutations does not guarantee developing cancer disease, but the mutation can increase the risk of getting those cancers.

In the field of medicine, these cancer types are classified as Hereditary Breast and Ovarian Cancer Syndrome (HBOC). The average life expectancy of individuals with BRCA1, without any interventions, is approximately 4.2 years shorter than that of non-carriers of the BRCA1 gene [4]. Significant advancements have been made in the medical and statistical modeling of breast cancer risk among BRCA1/2 carriers. These include the application of competing risk survival analysis based on breast and ovarian cancer outcomes developed by Choi et al. [5], as well as various clinical trials investigating risk-reducing treatment approaches for breast cancer patients [6]. Despite these efforts, it remains crucial to ensure statistical validity across these studies especially when missing data exists.

From a statistical perspective, the study is centered on a specific disease, which may introduce selection bias due to the sampling process. This bias arises from the selection criteria based on specific probands in each family. To mitigate this sampling bias, an ascertainment correction should be applied to the likelihood calculation, conditioning on the proband information. To accurately capture the heterogeneity between families in the context of time-to-cancer outcomes, the use of a frailty model is recommended. There are various choices for frailty distributions in survival analysis, including the Gamma distribution and the log-Normal distribution.

## 1.2 Motivation

Although numerous studies on risk assessment in susceptible populations and statistical advancements in dynamic prediction have significantly contributed to understanding BRCA1/2 families, the issue of missing data remains a substantial challenge, particularly in the context of survival outcomes. Over the past decade, several methodologies have been proposed to address missing data, including the Expectation-Maximization (EM) algorithm, the Monte-Carlo EM algorithm for cases where the E-step lacks a closed form, and Multiple Imputation (MI). However, when applying frailty models, which incorporate random effects in survival analysis, the literature addressing missing data is relatively sparse.

In genetic epidemiology, research is typically conducted on a family-wise basis. Therefore, considering the family structure when addressing statistical problems is both essential and unavoidable. Moreover, existing techniques for handling missing data must be carefully adapted, as the clustered nature of the dataset introduces additional complexity. Within the genetic framework, many variables, such as genetic information and polygenic risk scores (PRS), are not independent between individuals. Traditional methodologies often fail to account for family correlations and ascertainment bias. Given that families are selected based on a proband, it is crucial to apply ascertainment correction to minimize the selection bias. This situation presents an opportunity to further investigate and develop adequate methods for handling missing data, taking into account family correlations and ascertainment bias.

In this project, we aim to investigate the current implementation of Multiple Imputation (MI) methods for frailty models. Additionally, we propose a novel MI method that explicitly incorporates the kinship matrix during the imputation of genetically related variables. This proposed method will be evaluated by comparing it to existing MI methods and Complete Case Analysis (CCA).

## 1.3 Objectives

With the proposed MI method and the BRCA1 data, the objectives of this thesis are designed as follows:

1. To adapt the kinship correlations into the imputation step
2. To incorporate the ascertainment correction into the likelihood while considering that not all probands are affected
3. To assess the novel MI method via the calculation of the estimations, biases, and precisions through the simulation study
4. To apply the novel MI method and adjusted likelihood to model the BRCA1 family data

## **1.4 Organizations of the Thesis**

## Chapter 2

# Literature Reviews

### 2.1 Survival Analysis

Survival analysis is a robust statistical methodology used to analyze time-to-event data, where the focus is on the time until an event of interest occurs. It has been extensively applied in medical research, particularly in studies involving cancer, where events such as death or relapse are critical endpoints. The literature identifies several key methods and models that form the backbone of survival analysis. Kaplan and Meier [7] introduced the Kaplan-Meier estimator, a nonparametric statistic used to estimate survival functions from incomplete observations, which remains widely used due to its simplicity and effectiveness in handling censored data. Cox [8] proposed the proportional hazards model, which allows for the inclusion of covariates and has become a standard technique for assessing the effect of explanatory variables on survival. Collett [9], Machin et al. [10], and Kleinbaum and Klein [11] provide detailed expositions on survival analysis methods, including parametric and nonparametric approaches. Recent advancements have addressed complex issues such as interval censoring and competing risks, expanding the applicability and precision of survival analysis. These methodological developments have significantly enhanced the ability to make informed inferences from survival data, contributing to more accurate prognostic assessments and treatment evaluations in clinical research. In the survival analysis, there are several key functions that will contribute essentially to nearly all relevant scientific work.

The survival function  $S(t)$  measures the probability of study subject's entry age  $t$  is less than the event time  $T$ ,

$$S(t) = P(T > t) \tag{2.1}$$

The cumulative distribution function  $F(t)$  represents the probability that the event has occurred by time  $t$ . This is the complement of the survival function  $S(t)$ ,

$$F(t) = P(T \leq t) = 1 - S(t) \tag{2.2}$$

The probability density function  $f(t)$  is the likelihood of the event occurring at an exact time  $t$ . It is the derivative of the cumulative distribution function (CDF) or the negative derivative of the survival function.

$$f(t) = \frac{d}{dt}F(t) = -\frac{d}{dt}S(t) = h(t)S(t) \quad (2.3)$$

The hazard function  $h(t)$  measures the instantaneous rate at which the event occurs, given that the individual has survived up to the time  $t$ . It is the probability that an event occurs in a very small time interval, given survival until the beginning of the interval.

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} \quad (2.4)$$

The accumulated risk of experiencing the event up to time  $t$  can be expressed using the cumulative hazard function. It is the integral of the hazard function over time, providing a cumulative measure of risk.

$$H(t) = \int_0^t h(u)du = -\log S(t) \quad (2.5)$$

Whenever we can define one of above functions, it is straightforward to derive the rest.

### 2.1.1 Frailty Model for Family Based Study

Parametric survival analysis methods assume that the time-to-event data follow a specific probability distribution. This approach provides a more detailed and flexible framework for modeling survival data, allowing for more precise estimates and interpretations of survival functions and hazard rates. Hosmer Jr et al. [12] has comprehensively summarized most of parametric baseline hazard

# Chapter 3

## Methods

### 3.1 Introduction

When making the imputation on the continuous variable, one common way is to assume a conditionally normal distribution. This conditional distribution can be estimated from a linear regression. The original multiple imputation structure was brought by Rubin [13], that this method has been widely used by different scientific researchers with different models. There are many softwares that based on the multiple imputations as well, such as Blimp software [14], MICE package in R [15], and jomo package in R [16]. Apparently, this thesis cannot provide enough rooms for those well-designed softwares that I have not mentioned. The development of Multiple Imputation within certain specific models remains incomplete, and many details have not yet been comprehensively addressed in the current published literature. In genetic epidemiology, studies are mostly conducted with family clusters. Current implementation of the multiple imputation does not account for the genetic correlations. Furthermore, there remains room for exploration in the application of the frailty model with ascertainment correction. Therefore, this research is designed to develop a computationally efficient multiple imputation method to account for the kinship correlations, and apply it to frailty models with ascertainment correction.

In this chapter, a comprehensive guideline and adjusted multiple imputation formulas are provided. We explicitly show how the kinship matrix works in the imputation step, and how the ascertainment correction handles the analysis step in this research. The special imputation model is introduced as well, while considering the genetic variance and residual variance (sometimes refer to the environmental variance). The variance estimation using Rubin's Rule is provided, as well as the confidence interval based on completed data following the proposed multiple imputation.

### 3.2 Kinship Matrix

# Bibliography

- [1] JA Duncan, JR Reeves, and TG Cooke. BRCA1 and BRCA2 proteins: roles in health and disease. *Molecular pathology*, 51(5):237, 1998.
- [2] Bruce G Haffty, Elizabeth Harrold, Atif J Khan, Pradip Pathare, Tanya E Smith, Bruce C Turner, Peter M Glazer, Barbara Ward, Daryl Carter, Ellen Matloff, et al. Outcome of conservatively managed early-onset breast cancer by BRCA1/2 status. *The Lancet*, 359(9316):1471–1477, 2002.
- [3] Yong-Wen Huang. Association of BRCA1/2 mutations with ovarian cancer prognosis: an updated meta-analysis. *Medicine*, 97(2), 2018.
- [4] Michael P Lux, Peter A Fasching, and Matthias W Beckmann. Hereditary breast and ovarian cancer: review and future perspectives. *Journal of molecular medicine*, 84:16–28, 2006.
- [5] Yun-Hee Choi, Hae Jung, Saundra Buys, Mary Daly, Esther M John, John Hopper, Irene Andrulis, Mary Beth Terry, and Laurent Briollais. A competing risks model with binary time varying covariates for estimation of breast cancer risks in brca1 families. *Statistical Methods in Medical Research*, 30(9):2165–2183, 2021.
- [6] Yun-Hee Choi, Mary Beth Terry, Mary B Daly, Robert J MacInnis, John L Hopper, Sarah Colonna, Saundra S Buys, Irene L Andrulis, Esther M John, Allison W Kurian, et al. Association of risk-reducing salpingo-oophorectomy with breast cancer risk in women with brca1 and brca2 pathogenic variants. *JAMA oncology*, 7(4):585–592, 2021.
- [7] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.
- [8] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.
- [9] David Collett. *Modelling survival data in medical research*. Chapman and Hall/CRC, 2023.
- [10] David Machin, Yin Bun Cheung, and Mahesh Parmar. *Survival analysis: a practical approach*. John Wiley & Sons, 2006.



- [11] David G Kleinbaum and Mitchel Klein. *Survival analysis a self-learning text*. Springer, 1996.
- [12] David W Hosmer Jr, Stanley Lemeshow, and Susanne May. *Applied survival analysis: regression modeling of time-to-event data*, volume 618. John Wiley & Sons, 2008.
- [13] Donald B Rubin. Multiple imputation for survey nonresponse, 1987.
- [14] BT Keller and CK Enders. Blimp user’s guide (version 3). *Blimp Software: Los Angeles, CA, USA*, 2021.
- [15] Patrick Royston and Ian R White. Multiple imputation by chained equations (mice): implementation in stata. *Journal of statistical software*, 45:1–20, 2011.
- [16] Matteo Quartagno, Simon Grund, and James Carpenter. Jomo: a flexible package for two-level joint modelling multiple imputation. *R Journal*, 9(1), 2019.