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

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 upbringrings.
To my father, Guangmin Bi, for teaching me not to give up.
To my supervisors, for their unstopping guidance and invaluable lessons.

Acknowledgements

Contents

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
δ_{ij}	Event indicator for individual i in family j
w	The observed survival data (t, δ)
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
ω	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
$\mathscr{L}(\cdot)$	Laplace transform
$\mathbf{x}$	Covariates
$\boldsymbol{\beta}$	Model coefficients vector
$\boldsymbol{\theta}$	Parameter vector
Λ	The combination of $(\boldsymbol{\beta}, \lambda, \alpha)$

λ	Weibull shape parameter
α	Weibull scale parameter
υ	General form of the parameter in an undefined frailty distribution
k	Gamma shape and rate parameters
σ^2	Log-Normal variance parameter
ψ	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, 4]. Inheriting one of these mutations does not guarantee developing cancer disease, but the mutation can increase the risk of getting those cancers [5].

In the field of medicine, these cancer types are classified as Hereditary Breast and Ovarian Cancer Syndrome (HBOC) [6]. 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 [7]. 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. [8], as well as various clinical trials investigating risk-reducing treatment approaches for breast cancer patients [9]. 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

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] Julia B Greer and David C Whitcomb. Role of BRCA1/2 mutations in pancreatic cancer. *Gut*, 2006.

[3] 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.

[4] Yong-Wen Huang. Association of BRCA1/2 mutations with ovarian cancer prognosis: an updated meta-analysis. *Medicine*, 97(2), 2018.

[5] Bernard Friedenson. The BRCA1/2 pathway prevents hematologic cancers in addition to breast and ovarian cancers. *BMC cancer*, 7(1):1–11, 2007.

[6] 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.

[7] Phuong L Mai, Nilanjan Chatterjee, Patricia Hartge, Margaret Tucker, Lawrence Brody, Jeffery P Struewing, and Sholom Wacholder. Potential excess mortality in brca1/2 mutation carriers beyond breast, ovarian, prostate, and pancreatic cancers, and melanoma. *PLoS One*, 4(3):e4812, 2009.

[8] 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.

[9] 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.