

Design and Analysis of Studies Assessing Exposure Effects in Complex Settings

by

Kecheng Li

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics (Biostatistics)

Waterloo, Ontario, Canada, 2024

© Kecheng Li 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Liquun Wang
 Professor
 Department of Statistics
 University of Manitoba

Supervisor: Richard J. Cook
 University Professor
 Department of Statistics and Actuarial Science
 University of Waterloo

Internal Members: Glen McGee
 Assistant Professor
 Department of Statistics and Actuarial Science
 University of Waterloo

 Leilei Zeng
 Professor
 Department of Statistics and Actuarial Science
 University of Waterloo

Internal-External Member: Abel Torres Espin
 Assistant Professor
 School of Public Health Sciences
 University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis research has led to the following published article, submitted papers, and draft manuscript.

(a) Chapter 2 is based on

- (i) [Li et al \(2024a\)](#): Li K, Akkaya-Hocagil T, Cook RJ, Ryan LM, Carter RC, Dang KD, Jacobson JL, Jacobson SW. Use of generalized propensity scores for assessing effects of multiple exposures. *Statistics in Biosciences*. 2024 Jul;16(2):347-76.

(b) Chapter 3 is based on

- (i) [Li and Cook \(2024b\)](#): The design of large-scale plasma donation trials. *Submitted*.
- (ii) [Li and Cook \(2024c\)](#): Group sequential analysis of marked point processes: plasma donation trials. *Submitted*.

(c) Chapter 4 is based on

- (i) [Li and Cook \(2024d\)](#): Two-phase design of biomarker studies with progressive disease processes under intermittent observation. *Manuscript*.

All the papers are co-authored with my supervisor, and my contributions include formulating the research problems, deriving the main results, performing simulations in R, conducting real data analyses, and drafting the initial manuscripts.

Abstract

Understanding and efficiently estimating the effects of exposures on health outcomes is a fundamental goal in public health research. The accuracy and efficiency of exposure effect estimates are heavily influenced by study design and the statistical methodologies employed for analyses. This thesis consists of three projects where statistical methods are developed to address unique challenges in the estimation of exposure effects in diverse and complex settings.

The first project concerns causal inference regarding the effects of multiple exposure variables and their interactions which require modeling the joint distribution of exposures given pertinent confounding variables. This modeling can be carried out via a second-order regression model. Chapter 2 presents methodologies using regression adjustment and inverse weighting to investigate the asymptotic bias of estimators when the dependence model for the generalized propensity score incorrectly assumes conditional independence of exposures or is based on a naive dependence model which does not accommodate the effect of confounders on the conditional association of exposures. We also consider the problem of a semi-continuous bivariate exposure that arises when the population is made up of a sub-population of unexposed individuals and a sub-population of exposed individuals in whom the level of exposure can be quantified. We propose a two-stage estimation technique to study the effects of prenatal alcohol exposure, and specifically the effects of drinking frequency and intensity on childhood cognition in Chapter 2.

The second project focuses on plasma donation, highlighting the importance of rigorous device safety evaluations for donors. Complications arise from the fact that outcomes on successive donations from the same donor are not independent, adverse event rates are extremely rare, and there is substantial heterogeneity in the propensity for donors to donate over time. Chapter 3 introduces a statistical framework for designing superiority and non-inferiority trials to assess the safety of a new donation device compared to the standard one. A unique feature is that the number of donations per donor varies substantially so some individuals contribute more information and others less. Historical data on the donation rate and variation in the donation rate (heterogeneity) across donors, the adverse event rate, and the serial dependence in adverse events provide the necessary information to plan for the duration of accrual and follow-up periods. Specifically the sample size formula is derived to ensure power requirements are met when analyses are based on generalized estimating equations and robust variance estimation. The complexity of recruiting donors from a heterogeneous and dynamic population of donors means that it is challenging to characterize the rate at which information is acquired on treatment effects over the course of the study. Strategies for interim monitoring based on group sequential designs using

alpha spending functions are developed based on a robust covariance matrix for estimates of treatment effect over successive analyses. The design of a plasma donation study is illustrated in Chapter 3 aiming to investigate the safety of a new device with the outcome being serious hypotensive adverse events.

Many chronic diseases can be naturally characterized using multistate models. Longitudinal cohorts and registry studies of chronic diseases typically recruit and follow individuals to record data on the nature and timing of disease progression. In many cases the exact transition times between disease states are not observed directly, but the state occupied at each clinic visit is known. Such studies also routinely collect and store serum samples at intermittent clinic visits. The final project explores the design and analysis of two-phase studies for evaluating the effect of a biomarker of interest. We consider the design of two-phase studies in Chapter 4 aimed at selecting individuals for biospecimens assays to measure biomarkers of interest and estimate their association with disease progression through intensity-based modeling. Likelihood-based and estimating function approaches are developed and the efficiency gains from score residual-dependent sampling strategies are investigated for joint models of the biomarker and disease progression processes. The efficiency of these different frameworks is investigated, and the methods are applied to a study investigating the association between the HLA-B27 marker and joint damage in psoriatic arthritis patients.

The thesis concludes with some topics of future research in Chapter 5.

Acknowledgements

I would like to express my gratitude to all those who have supported and guided me throughout my PhD journey. I wish to give my deepest and heartfelt thanks to my supervisor, Dr. Richard J. Cook, for his invaluable guidance, patience, and unwavering support over the past few years. His expertise, encouragement, and insightful feedback have been instrumental in the completion of this thesis. I feel incredibly fortunate to have the best supervisor during the PhD journey in my life.

I would like to extend my sincere thanks to the members of my thesis committee, Dr. Liqun Wang, Dr. Glen McGee, Dr. Leilei Zeng, and Dr. Abel Torres Espin, for their time and valuable suggestions throughout the research process. I would also like to thank our wonderful Department of Statistics and Actuarial Science for providing a highly supportive academic environment with great resources. Special thanks go to Ms. Mary Lou Dufton and Mr. Greg Preston whose assistance has made the logistics of this journey much smoother.

I would also like to thank Dr. Louise Ryan, Dr. Tugba Akkaya Hocagil, and Dr. Doan Khue Dung for insightful discussions regarding the prenatal alcohol study, and Dr. Joseph Jacobson and Dr. Sandra Jacobson for granting permission to use data from the Detroit Study. I am also thankful to Professor Nancy Heddle and Dr. Sheila O'Brien for discussions on the design of platelet donation trials. Thanks also go to Dr. Dafna Gladman, Dr. Vinod Chandra, Dr. Lihi Eder, and Mr. Daniel Pereira for discussions regarding the research program at the University of Toronto Centre for Prognosis Studies in Rheumatic Diseases, as well as for permission to use the data from the University of Toronto Psoriatic Arthritis Clinic.

To my friends in Canada and China, Xianwei, Alexandra, Tong, Laura, Pusheng, Yuan, Xiyue, Yechao, Chi, Mu, Yimiao, Jingyue, Ziqian, Jianchu, Zhaoran, Liyuan, Bingying, Yiran, Mei, Chi-kuang, Yiming, Peijun, Qinglong, Sinan, Xiaoya, Aelita, Lily, Jiayi, Ann, Dingding, Zijin, Henan, Yuliang, Tiancheng, Yuying, Fangya, Zhaohan, Ce, Yishu, Linke, Zhongyuan, Jingwen, Yuwen, Xin, Xisi, Wei, Meiyi, Jili, Ziyang, Mingguang, Kechen, Jiaqin and Ye. Thank you for all the joyful moments we shared, especially in Mathematics 3.

Lastly I am deeply grateful to my family. To my parents Peihong and Hengping, and my partner Hwanwoo, thank you for believing in me and providing me with the courage to pursue my dreams. Thank you for the endless understanding and patience which kept me motivated through all the ups and downs of this journey. Thank you for loving me.

Dedication

To my family.

Table of Contents

Examining Committee	ii
Author's Declaration	iii
Statement of Contributions	iv
Abstract	v
Acknowledgements	vii
Dedication	viii
List of Figures	xiv
List of Tables	xvii
1 Introduction	1
1.1 Overview	1
1.2 Generalized Propensity Scores	2
1.2.1 Causal Inference with Multiple Exposures	2
1.2.2 Prenatal Alcohol Exposure and Childhood Cognition	4
1.3 Interim Analysis of Marked Point Processes	4
1.3.1 Group Sequential Analyses	4

1.3.2	The IMPACT Trial in Blood Donation Research	7
1.4	Biomarker Study of Progressive Disease Processes	7
1.4.1	Two-Phase Designs	7
1.4.2	Joint Damage in Psoriatic Arthritis	8
1.5	Outline of Thesis	9
2	Generalized Propensity Scores for Assessing Effects of Multiple Exposures	11
2.1	Introduction	11
2.2	Causal Analysis with Multiple Exposures	12
2.2.1	Specification of the Full Model	12
2.2.2	Propensity Scores for Two Binary Exposures	13
2.3	Exposure Dependence Model Misspecification	16
2.3.1	Regression Adjustment	16
2.3.2	Inverse Probability Weighting	18
2.3.3	Empirical Bias and Efficiency	19
2.4	Multivariate Continuous Exposures	20
2.4.1	Multivariate Normal Exposures	23
2.4.2	Multivariate Semi-continuous Exposures	24
2.4.3	Misspecification of Propensity Scores	26
2.5	Two-Stage Causal Analysis for Bivariate Semi-Continuous Exposures	29
2.5.1	Stage I: The Dose-Response Surface	31
2.5.2	Stage II: The Effect of Exposure Status	31
2.5.3	Simulation Studies on Two-Stage Approach	34
2.6	Prenatal Alcohol Exposure and Cognition	36
2.6.1	Subgroup Analysis involving Drinkers	36
2.6.2	Analyses based on Bivariate Semi-Continuous Exposure Model	38
2.6.3	Full Sample Analysis with Two-Stage Approach	38
2.6.4	Results	38
2.7	Remarks	41

3	The Design of Large-Scale Plasma Donation Trials	43
3.1	Introduction	43
3.2	Statistical Formulation and Methods	45
3.2.1	Notation and Generalized Estimating Equations	45
3.2.2	Large Sample Results for Study Design	47
3.3	A Simplified Setting for Trial Design	48
3.3.1	Large Sample Results	48
3.3.2	Zero-Truncated Conditional Negative Binomial Process for Donations	50
3.4	Study Design, Validation and an Application	54
3.4.1	Sample Size for Superiority and Non-inferiority Trials	54
3.4.2	Empirical Studies	56
3.4.3	Design of a Plasma Donation Study	60
3.5	Remarks	64
3.6	Sequential Analysis of Marked Point Processes	65
3.7	Statistical Formulation and Methods	67
3.7.1	Notation and Generalized Estimating Equations	67
3.7.2	Joint Distribution of Sequential Test Statistics	71
3.7.3	The Independent Increments Structure	73
3.8	Results for a Simple Two-Arm Trial	74
3.8.1	Derivations for Matrices $\mathcal{A}(t_r)$, $\mathcal{B}(t_r)$ and $\mathcal{B}(t_r, t_s)$	74
3.8.2	Derivations for $M_1(t_r)$, $M_2(t_r, t_s)$ and $M_3(t_r)$	75
3.8.3	The Asymptotic Variance of $\hat{\beta}_1(t_r)$	76
3.9	Modelling Accrual Time and Donation Rates	77
3.10	Sample Size and Group Sequential Testing	79
3.10.1	Alpha Spending Functions	79
3.10.2	Sample Size for a Sequential Superiority Trial	80
3.10.3	Sample Size for a Sequential Non-Inferiority Trial	81

3.11	Simulation Studies	82
3.11.1	Empirical Validation with Sequential Trials	82
3.11.2	Design of a Plasma Donation Study	85
3.12	Remarks	90
4	Two-Phase Designs for Biomarker Studies on Disease Progression	91
4.1	Introduction	91
4.2	Intermittent Observation of Progressive Processes	94
4.3	Analysis Methods for Two-Phase Studies	96
4.3.1	Maximum Likelihood	96
4.3.2	Conditional Maximum Likelihood	99
4.3.3	Inverse Probability Weighted Estimating Equations	100
4.4	Two-Phase Sub-Sampling Strategies	101
4.4.1	Residual Dependent Phase II Subsampling	101
4.4.2	Stratified Phase II Subsampling	103
4.5	Empirical Studies of Two-Phase Designs	104
4.6	Joint Damage in Psoriatic Arthritis	105
4.7	Remarks	109
5	Review and Future Work	111
5.1	Confounding with Multivariate Exposures	111
5.2	Design of Plasma Donation Trials	112
5.3	Two-Phase Designs with Complex Diseases	113
	References	115
	APPENDICES	131

A	Appendix to Chapter 2	132
A.1	Impact of a Misspecified Generalized Propensity Score	132
A.2	Second-Order Generalized Estimating Equations (GEE2)	135
A.3	Covariance of Causal Effects of Multiple Exposures	138
A.4	Supplementary Figures and Tables	141
B	Appendix to Chapter 3	146
B.1	Derivations of Asymptotic Variance of $\hat{\beta}_1$ in a Simplified Setting	146
B.2	Conditional Mean and Variance of $K_F \mid K_A \geq 1$	150
B.3	Conditional Mean and Variance of $K_F \mid L \leq \tau_A$	153
B.4	Matrices $\mathcal{A}(t_r)$, $\mathcal{B}(t_r)$ and $\mathcal{B}(t_r, t_s)$ in a Simplified Setting	155
B.5	Derivations for Design of a Two-Arm Clinical Trial	163
B.6	Derivations of $\kappa_x(t_r)$, $\nu_x(t_r)$, and $\kappa_x(t_r, t_s)$	165
B.7	Supplementary Table of One-sided Boundaries	171
C	Appendix to Chapter 4	173
C.1	Derivations of Score Function	173
C.2	Derivations of Score-Type Residuals M_μ	174
C.3	Supplementary Table of Strata Sizes	176

List of Figures

1.1	An illustration of the distribution of a bivariate semi-continuous exposure.	3
1.2	An illustration of the recruitment process and repeated measurements in a randomized controlled trial with planned interim analyses at study times t_1 , t_2 and t_3 .	5
1.3	Disease progression in terms of the number of damaged joints for patients with psoriatic arthritis.	9
2.1	Empirical relative bias with standard errors (shaded bands) of estimates of exposure effects from inadequate propensity score modeling assuming conditional independence between two binary exposures when the second-order confounder has the effect γ_{31} ; regression adjustment (Reg) (yellow+red) and inverse weighting (IPW) (blue) are considered; the marginal association between two binary exposures is $\mathbb{E}\{\text{OR}(A_1, A_2)\} = 4$.	21
2.2	Empirical relative bias with standard errors (shaded bands) of estimates of exposure effects from inadequate propensity score modeling due to the omission of a second-order confounder with effect γ_{31} ; regression adjustment (Reg) (red) and inverse weighting (IPW) (blue) are considered; the marginal association between two binary exposures is $\mathbb{E}\{\text{OR}(A_1, A_2)\} = 4$.	22
2.3	Empirical relative bias with standard errors (shaded bands) of estimates of exposure effects from inadequate propensity score modeling assuming conditional independence between two continuous exposures when the second-order confounder has the effect γ_{31} ; regression adjustment (Reg) (red) and inverse weighting (IPW) (blue) are considered; the marginal association between two continuous exposures is $\mathbb{E}(\rho) = 0.5$.	28

2.4	Empirical relative bias with standard errors (shaded bands) of estimates of exposure effects from inadequate propensity score modeling due to the omission of a second-order confounder with effect γ_{31} ; regression adjustment (Reg) (red) and inverse weighting (IPW) (blue) are considered; the marginal association between two continuous exposures is $\mathbb{E}(\rho) = 0.5$	30
2.5	Dose-response surface and contour plot generated from stage I in two-stage analysis: the effect of changes in drinking frequency (proportion days drinking) and intensity (dose/occasion) on child cognitive deficit; reference levels: 1.81 oz AA/occasion, 0.13 of the pregnancy drinking.	41
3.1	An illustration of 10 individuals recruitment process and repeated measurements.	51
3.2	Impact of donor population size (N_{pop}) on the number of donors required for recruitment ($\mathbb{E}\{N(\tau_A)\}$), average number of donations per donor (κ), total number of donations to be collected ($\kappa\mathbb{E}\{N(\tau_A)\}$), and recruitment duration (τ_A) in a non-inferiority trial with $\lambda = 21$, $\mu_0 = 0.035\%$, $\exp(\beta_{1A}) = 1.5$, $\exp(\beta_{1M}) = 2.5$, $\rho = 0.005$, $\phi = 0.8$, 1.6 , $\tau_F = 0.5$, $\alpha = 5\%$, target power=80%, and consent rate=60%, 100%.	61
3.3	Power of a non-inferiority trial when $\exp(\beta_{1A})$ and ρ are misspecified; true parameters are set to $N = 40,000$, $\lambda = 21$, $\mu_0 = 0.035\%$, $\exp(\beta_{1A}) = 1.5$, $\exp(\beta_{1M}) = 2.5$, $\rho = 0.005$, $\phi = 0.8$, $\tau_F = 0.5$, $\alpha = 5\%$, and target power=80%.	63
3.4	An illustration of the recruitment process for 10 subjects along with their repeated measurements. Interim analyses are planned at study times $t_1, t_2, \dots, t_{R-1}$, and a final analysis at t_R . The length of the accrual period is τ_A , the length of the follow-up period is τ_F for each individual, and the entire study duration is $\tau = \tau_A + \tau_F$	68
3.5	Study progression over time in a non-inferiority trial of information fraction, number of donors recruited, average donations per donor, and total donations with $\lambda = 21$, $\phi = 0.8$, $\mu_0 = 0.035\%$, $\exp(\beta_{1A}) = 1.5$, $\exp(\beta_{1M}) = 2.5$, $\rho = 0.005$, $\tau_F = 0.5$, $\alpha = 5\%$, and target power=80%.	89
4.1	Ten sample timelines depicting disease progression obtained from follow-up visits since clinic entry on joint damage in patients with psoriatic arthritis (PsA) from the University of Toronto Psoriatic Arthritis Cohort.	93
4.2	A progressive multistate model with $K + 1$ states.	95

4.3	A progressive model for the development of damaged joints in patients with psoriatic arthritis.	107
5.1	An illustration of three disease processes.	113
A.1	Empirical relative bias with standard errors (shaded bands) of estimates of exposure effects from inadequate propensity score modeling assuming conditional independence between two semi-continuous exposures when the second-order confounder has the effect γ_{31} ; regression adjustment (Reg) (red) and inverse weighting (IPW) (blue) are considered; the marginal association between two semi-continuous exposures is $\mathbb{E}(\rho) = 0.5$	144
A.2	Empirical relative bias with standard errors (shaded bands) of estimates of exposure effects from inadequate propensity score modeling due to the omission of a second-order confounder with effect γ_{31} ; regression adjustment (Reg) (red) and inverse weighting (IPW) (blue) are considered; the marginal association between two semi-continuous exposures is $\mathbb{E}(\rho) = 0.5$	145

List of Tables

2.1	Simulation results with sample size $n = 500$, repetitions $n_{sim} = 1000$: relative bias and standard error of stage II estimator of the binary exposure effect when models are correctly specified.	35
2.2	Simulation results with sample size $n = 500$, repetitions $n_{sim} = 1000$: relative bias and standard error of stage II estimator of the binary exposure effect with model misspecification in stage II.	35
2.3	Estimates of the effects for each confounder.	37
2.4	Estimated effects of being a drinker versus non-drinker, the logarithm of prenatal absolute alcohol exposure per occasion, the logarithm of proportion of drinking days across pregnancy, and their interaction.	39
3.1	Length of recruitment duration (τ_A) and expect number of donors ($\mathbb{E}(N)$) required across sites for a superiority trial with $\mu_0 = 0.002$, $\exp(\beta_{1A}) = 1.5$, $\tau_F = 0.5$, and $\beta_1^\circ = 0$; verification by calculation of the empirical type I error and the empirical power; $n_{sim} = 1000$	58
3.2	Length of recruitment duration (τ_A) and expect number of donors ($\mathbb{E}(N)$) required across sites for a non-inferiority trial with $\mu_0 = 0.002$, $\exp(\beta_{1A}) = 1.5$, $\tau_F = 0.5$, $\beta_1^\circ = 0$, and non-inferiority margin $\exp(\beta_{1M}) = 2$; verification by calculation of the empirical type I error and the empirical power; $n_{sim} = 1000$	59
3.3	Sample size required in terms of the number of subjects needed for a superiority design with $\mu_0 = 0.002$, $\exp(\beta_{1A}) = 1.2$, $\tau_A = 0.5$, $\tau_F = 1$, and $\beta_1^\circ = 0$; two interim analyses are planned with $R = 3$	83

3.4	Sample size required in terms of the number of subjects needed for a non-inferiority design with $\mu_0 = 0.002$, $\exp(\beta_{1A}) = 1.2$, $\tau_A = 0.5$, $\tau_F = 1$, and non-inferiority margin $\exp(\beta_{1M}) = 1.5$; two interim analyses are planned with $R = 3$	84
3.5	Empirical rejection rates for a superiority trial with two interim analyses. .	86
3.6	Empirical rejection rates for a non-inferiority trial with two interim analyses.	87
4.1	Tabulation of strata in empirical studies of two-phase design.	105
4.2	Empirical standard errors and analytical standard errors [†] of $\hat{\beta}_1$ obtained from maximum likelihood (ML), conditional maximum likelihood (CML), and inverse probability weighting (IPW) approaches for two-phase designs with simple random sampling (SRS), balanced sampling (BAL), and score residual-based sampling (RES); phase I sample size $n = 2000$, $n_{sim} = 1000$.	106
4.3	Estimates with standard errors from maximum likelihood (ML), conditional maximum likelihood (CML), and inverse probability weighting (IPW) approaches for the two-phase designs with simple random sampling (SRS), balanced sampling (BAL), and residual-based sampling (RES) employed in the joint damage data of patients with psoriatic arthritis; phase I sample size $n = 1116$, and phase II sample size $n_2 = 480$	108
4.4	Summary of estimates obtained from maximum likelihood (ML) with residual-dependent sampling (RES) employed in phase I; covariates include HLA-B27 status, gender, age of onset of psoriatic arthritis (years), and family history of psoriatic arthritis.	108
A.1	Summary of characteristics of subjects in Detroit cohort study of prenatal alcohol exposure (PAE) and children cognition.	141
A.2	Summary of covariate balance in terms of maximum absolute correlation and average absolute correlation.	143
B.1	One-sided boundaries from different alpha spending functions for a non-inferiority trial with 1st interim analysis planned at study time t_1 , 2nd interim analysis planned at study time t_2 , and final analysis planned at the end of study; information fraction is determined as the fraction of inverse variance of $\hat{\beta}_1$; $\lambda = 21$, $\phi = 0.8$, $\mu_0 = 0.035\%$, $\exp(\beta_{1A}) = 1.5$, $\exp(\beta_{1M}) = 2.5$, $\rho = 0.005$, $\tau_A = 0.5$, $\tau_F = 0.5$, $\alpha = 5\%$, and target power=80%.	172

C.1	Average strata sizes based on 1000 replicates [†]	177
-----	--	-----

Chapter 1

Introduction

1.1 Overview

One of the primary objectives in public health research is to understand and estimate exposure effects on outcomes of interest. The validity and efficiency of exposure effect estimation are influenced by the type of study design and accompanying analysis. This thesis considers design and analysis methodology addressing three challenges in broad areas of research:

- (i) using generalized propensity scores for causal analysis with multiple exposures,
- (ii) evaluating intervention effects administered repeatedly in randomized trials through repeated significant testing with interim analyses, and
- (iii) designing two-phase studies for the cost-effective assessment of a biomarker effect on the disease process which is under intermittent observation.

In Section [1.2](#) to [1.4](#) a brief overview is provided of the statistical methods developed in this thesis and three motivating problems. The details regarding the statistical research are developed and discussed in Chapters [2](#) to [4](#) respectively.

1.2 Generalized Propensity Scores

1.2.1 Causal Inference with Multiple Exposures

Estimation of exposure effects is a common and scientifically important aim in many epidemiology studies, but it is challenging in observational research due to the complex nature of the observational data and the effect of confounding. [Rosenbaum and Rubin \(1983\)](#) introduced the idea and described the utility of propensity scores for causal analysis in order to mitigate the impact of confounding variables. The propensity score was initially presented in the context of a binary exposure variable in which case it is defined as the conditional probability of being exposed given a set of covariates. Through matching ([Rosenbaum and Rubin, 1983, 1985](#)), stratification ([Rosenbaum and Rubin, 1984](#); [Cochran, 1968](#)), regression adjustment ([Haukoos and Lewis, 2015](#)) and inverse probability weighting (IPW) ([Rosenbaum, 1987](#); [Lunceford and Davidian, 2004](#)) the estimated propensity score can be shown to mitigate the effect of confounding variables. In settings where randomized controlled trials (RCTs) are impractical or unethical this enables researchers across various disciplines (e.g. social sciences, economics and epidemiology) to account for the imbalance in the observed characteristics across treated and untreated individuals to improve the assessment of causal effects. Throughout the thesis the words treatment and exposure are used interchangeably.

The concept of a generalized propensity score was developed to deal with settings where the exposure variable is continuous ([Imbens, 2000](#); [Hirano and Imbens, 2004](#)). For example we might be interested in the dosage of a new drug on reducing the risk of hypertension. Meanwhile the complex etiology of many chronic conditions motivates joint consideration of multiple exposure variables and their possible interacting effects ([Dominici et al, 2010](#)). Examples include the study of smoking and alcohol consumption on pancreatic cancer ([Duell, 2012](#)), the interacting effect of smoking and drinking on dietary intake of antioxidants on oral, pharyngeal or laryngeal squamous cell carcinoma ([Suzuki et al, 2006](#)), and environmental studies of several different air pollutants and their effect on lung cancer risk ([Beelen et al, 2008](#)). Oftentimes the exposure variable is semi-continuous, meaning that the data includes a substantial proportion of individuals with no exposure at all ([Akkaya Hocagil et al, 2021](#)). Figure 1.1 shows an example of a bivariate semi-continuous exposure from our motivating study in Section 1.2.2 where the prenatal alcohol drinking intensity (x-axis) and drinking frequency (y-axis) for pregnant women in the study cohort are presented. For mothers with no prenatal alcohol consumption, both drinking intensity and frequency are zero, which leads to the zero-inflated exposure variables in the study data. We are then interested in assessing the effect of the number of drinking occasions per month (drinking

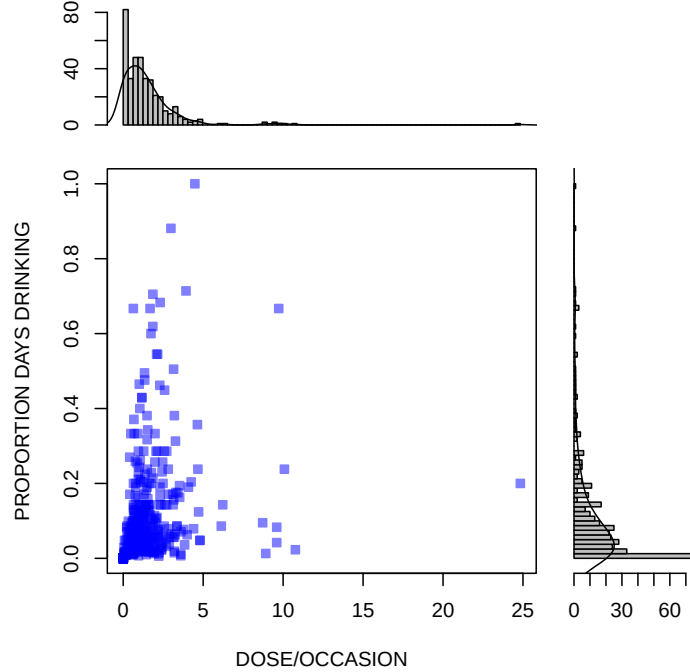


Figure 1.1: An illustration of the distribution of a bivariate semi-continuous exposure.

frequency), daily ounces of alcohol intake (drinking intensity), and the interaction between these factors on their children’s cognitive outcomes. When the joint effects of multiple exposures are under consideration, multivariate propensity scores can play a useful role. [Vegetabile et al \(2020\)](#) describe a joint model for multivariate binary exposures through the specification of a model for a latent multidimensional Gaussian variable. [Williams and Crespi \(2020\)](#) develop a generalized propensity score and use it to construct inverse multivariate density weighted estimating functions. [Chen and Zhou \(2023\)](#) propose the use of entropy balancing for multiple continuous treatments as a framework for causal analysis; this approach is best suited for settings in which the dimensions of the exposure and confounding variables are modest. Based on prior work on the generalized propensity score we develop methods to evaluate the causal effects of multiple exposures and their interactions on the outcome of interest. We consider various types of exposures including binary, continuous, and semi-continuous forms. The implications of misspecifying the association between exposures are explored in [Chapter 2](#).

1.2.2 Prenatal Alcohol Exposure and Childhood Cognition

Fetal alcohol spectrum disorders (FASD) are the most common preventable cause of neurodevelopmental disabilities. The prevalence estimates of FASD is 2–5% among live births in North America (May et al, 2009), and the estimates can be as high as 20% among live births in endemic communities such as the Western Cape Province of South Africa (May et al, 2013). Fetal alcohol syndrome (FAS) is the most severe form of FASD and is characterized by distinctive craniofacial dysmorphism, microcephaly and growth retardation (Stratton et al, 1996; Hoyme et al, 2005). Milder variants include a) partial FAS (PFAS), diagnosed when there is facial dysmorphism, a history of prenatal alcohol exposure (PAE), and either growth retardation, microcephaly, or central nervous system (CNS) impairment, and b) alcohol-related neurodevelopmental disorder (ARND), diagnosed following PAE when there is cognitive and/or behavioural impairment but no facial dysmorphism. A broad range of cognitive and behavioral deficits occur in children with FASD including decreased IQ (Jacobson et al, 2004), lower verbal, learning, and memory skills (Lewis et al, 2015), and poorer executive function (Kodituwakku, 2007). Social problems, attention deficits, and aggressive behaviours have been noted (Carmichael Olson et al, 1998; Jacobson et al, 2006; Day et al, 2013), as has attention deficit hyperactivity disorder (Coles et al, 1997; Jacobson et al, 2011).

Even though previous studies show that high levels of prenatal alcohol exposure (PAE) can cause serious developmental problems in children (Jacobson et al, 2004; Kodituwakku, 2007; Lewis et al, 2015; Jacobson et al, 2011), the impact of low levels of exposure is not well-understood. Still less is known about the impact of “dose rate” which refers to the possibility that infants exposed to the same total volume of alcohol over the course of a pregnancy may have very different cognitive deficits depending on whether the mother drinks small amounts relatively frequently, or whether she drinks large amounts but less frequently (chronic versus binge drinking). This motivates the development and study of the methods in Chapter 2 to better understand possible complex causal effects of prenatal alcohol exposure on cognition outcomes of their children.

1.3 Interim Analysis of Marked Point Processes

1.3.1 Group Sequential Analyses

It is often ethically essential to conduct interim analyses during the course of large clinical trials to allow for the potential early termination and more rapid dissemination of study

results to individuals who can benefit from it. In the event of either a clear advantage or the emergence of harmful side effects, early termination should be considered. This is crucial not only to benefit future patients as soon as possible but, more importantly, to prevent the harmful exposure to others. In addition to ethical reasons, there are budgetary considerations for planning interim analyses as the costs associated with conducting a trial can be minimized for the possible early termination.

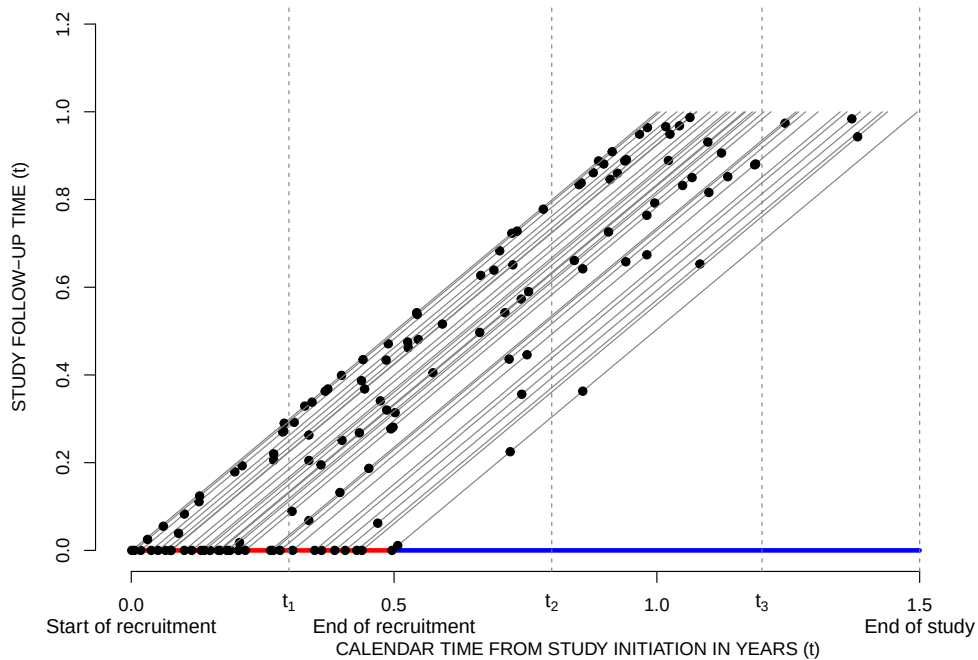


Figure 1.2: An illustration of the recruitment process and repeated measurements in a randomized controlled trial with planned interim analyses at study times t_1 , t_2 and t_3 .

When evaluating the exposure effect through repeated significance testing during the accumulation of data, researchers need to adjust for the inflation of the type I error to maintain the overall significance level ([Armitage et al, 1969](#)). [Pocock \(1977\)](#) suggested using one conservative critical value based on the work by [Armitage et al \(1969\)](#), as a means to control the type I error. Alternatively, methods proposed by [O'Brien and Fleming \(1979\)](#) and [Slud and Wei \(1982\)](#) provide a way to construct a different critical value for each interim analysis. These methods all require a predetermined number of interim analyses.

Lan and DeMets (1983) proposed the use of an alpha spending function which enables a more flexible way to compute the critical values. For studies where each subject is assessed with a single measurement, the sequentially computed test statistics would have an independent increment structure if we assume the independence between subjects. This independent increment structure allows for a simplified computation of critical values or group sequential boundaries for the test statistics by implementing recursive univariate integration (Armitage et al, 1969; Reboussin et al, 2000) using the approach of Lan and DeMets (1983). However in longitudinal studies where the responses of a subject are measured repeatedly over time, the independent increment structure of the response or test statistics may not hold due to the correlation within each subject. In such cases the method proposed by Lee and Demets (1991) based on random-effects models can be applied for continuous responses. Additionally, Wei et al (1990) and Lee et al (1996) have presented an approach based on generalized estimating equations (GEE) (Liang and Zeger, 1986) that is applicable to a broader range of data types.

In blood transfusion research there are two streams of research where information is acquired about the effect of an intervention according to a marked point process. In blood donation research individuals donate blood repeatedly over time and interest may lie in studying the occurrence of adverse events upon blood donation - in particular in comparing whether two methods of blood collection lead to similar adverse event rates. Figure 1.2 illustrates a donation study with an accrual period of six months (0.5 years) and one year of follow-up planned for each recruited donor. The horizontal axis records the time since the start of the study and the vertical axis records the time on study for recruited donors. The 45° lines emanating from the x-axis represent the passage of calendar time and study follow-up time for each recruited individual. The dots on the lines represent the times of donations, and the response of interest is recorded at such times. Three interim analyses are planned at calendar time t_1 , t_2 , and t_3 with a final analysis planned at the end of the study (represented by vertical dashed lines). The propensity to donate will vary across individuals and may also be associated with the time of recruitment. Our interest is in assessing the safety of using different blood collection devices through interim monitoring. The second setting where marked point processes arise in transfusion research is in the treatment of patients with anemia or thrombocytopenia and other chronic conditions. Here patients need for blood products (red blood cells and platelets respectively) will vary across individuals according to the severity of their underlying condition. Interest will often lie in comparing the therapeutic effect of different blood products (i.e. pathogen inactivated blood products versus standard blood products) on short-term outcomes in patients (i.e. hemoglobin or platelet levels). Here we focus our discussion primarily on the former transfusion setting in the blood donation study but the methods discussed are

equally motivated by and applicable to problems in transfusion medicine research.

1.3.2 The IMPACT Trial in Blood Donation Research

Source plasma is crucial in life-saving treatments for various medical conditions and the demand for it has consistently exceeded its supply, especially during the COVID-19 pandemic (Covington et al, 2023). For a potential increase in average plasma collection volume, the trial named IMPACT (IMproving Plasma CollecTion) (Hartmann et al, 2021) was conducted between January 6 and March 26, 2020 at three plasma collection centers (i.e. Charlotte, NC; Spokane, WA; Wichita, KS) in the United States. It is a randomized controlled trial to assess the non-inferiority of a novel percent plasma nomogram (PPN) compared to the current simplified 1992 nomogram in terms of the significant hypotensive adverse events among donors, with a predefined margin of 0.15%. A total of 3443 donors were enrolled across three sites and were randomly assigned to either the experimental arm utilizing the PPN with personalized target volume calculation for the plasma donation, or the control arm using the standard nomogram. The primary endpoint is the incidence rate of significant hypotensive adverse event which is defined based on the modified IQPP classification (see supporting information in Hartmann et al (2021)), and the secondary endpoint is the total plasma volume (TPV) collected in the study.

We consider the design of plasma donation trials in Chapter 3 where the goal is to assess the safety of a new device on the response to transfusions compared to the standard device. Strategies for interim monitoring based on group sequential designs using alpha spending functions (Lan and DeMets, 1983) are then developed based on a robust covariance matrix for estimates of treatment effect over successive analyses.

1.4 Biomarker Study of Progressive Disease Processes

1.4.1 Two-Phase Designs

In many modern cohort studies the biobanks have been established for the collection and storage of various biospecimens, including tissue samples and serum. However it might not be economically feasible to assess biospecimens for all individuals given the substantial costs associated with the biomarker assays. In cases where the interest is in the association between an expensive biomarker and the disease outcome, a two-phase study design provides a cost-efficient way to estimate the biomarker effect. The first phase (phase I)

of the two-phase study contains the inexpensive covariates and outcomes available for all subjects in the study. Subsequently, the second phase (phase II) focuses on the assessment of the expensive variable of interest, but this is only conducted for the selected subjects due to budgetary constraints.

The information in phase I can be used for the subsample selection in phase II. Many studies have an extreme-outcome sampling design, selecting subjects with the highest or lowest values of a continuous outcome; see [Chen and Li \(2011\)](#); [Li et al \(2011\)](#); [Lin et al \(2013\)](#). For failure time outcomes, [Ding et al \(2014\)](#) suggested selecting cases with the largest or smallest values of failure times and [Lawless \(2018\)](#) suggested to select the subjects with the smallest failure times or largest censored times. [Tao et al \(2020\)](#) introduced an efficient two-phase design for continuous, discrete or censored outcomes which selects the subjects with extreme values of a score-type residual. Challenges arise when our interest is in the effect of an expensive biomarker on the disease progression, especially when exact transition times between disease states are not directly observed. Oftentimes only the state occupied at each clinic visit is known, and biospecimens are also routinely collected and stored at these clinic visits. For this setting with intermittent observations [Kalbfleisch and Lawless \(1985\)](#) proposed a scoring procedure to obtain estimates under a time-homogeneous Markov model. We consider the design of two-phase studies in Chapter 4 where the objective is to select the biospecimens to assay in order to efficiently measure the biomarker effect on the disease progression.

1.4.2 Joint Damage in Psoriatic Arthritis

Chapter 4 is motivated by the aim to evaluate the biomarker effect associated with the development of joint damage in a two-phase study design. Psoriatic arthritis (PsA) is a chronic immunological disease where the immune system mistakenly attacks healthy joints leading to joint inflammation and potentially joint damage. The University of Toronto Psoriatic Arthritis Clinic was established in 1978 at the Centre for Prognosis Studies in Rheumatic Disease at the Toronto Western Hospital ([Gladman and Chandran, 2011](#)). Patients diagnosed with psoriatic arthritis were recruited in the PsA cohort and evaluated upon entry into the study, with subsequent radiographic examinations conducted every six months ([Gladman et al, 1987](#)). Patients had assessments across 64 joints of the hands and feet. One way to determine the disease states is based on the number of damaged joints where a damaged joint is identified as deformed, ankylosed, flail, or surgically treated. Based on the number of damaged joints, patients can be classified as having very mild disease (0 damaged joint), oligoarthritis (1-4 damaged joints), polyarthritis (5-9 damaged joints) and severe polyarthritis (> 9 damaged joints). The disease process regarding the

number of damaged joints is shown in Figure 1.3. We consider the design of two-phase studies in Chapter 4 aimed at selecting biospecimens for assay in order to efficiently measure the biomarker effect on the transition intensities in progressive disease processes under intermittent observation. The methods are applied to this motivating study to assess the relationship between the HLA-B27 marker and the progression of joint damage in PsA patients.

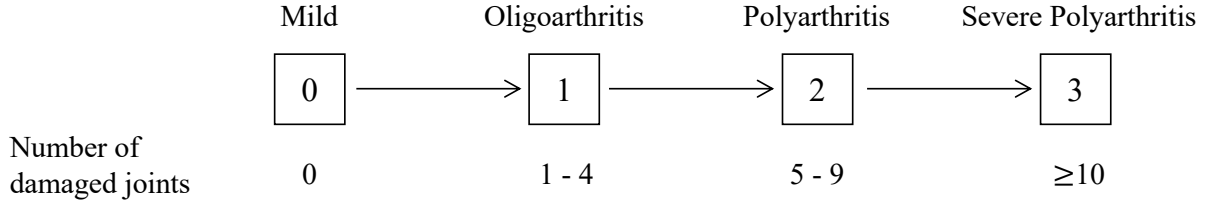


Figure 1.3: Disease progression in terms of the number of damaged joints for patients with psoriatic arthritis.

1.5 Outline of Thesis

The remainder of the thesis is organized as follows. In Chapter 2, our interest lies in causal analysis of the effects of multiple exposures and their interaction on an outcome. We use generalized propensity scores through the joint modeling of exposures for causal analysis via regression adjustment and inverse weighting. In both frameworks, we investigate the empirical bias and efficiency of estimators when the dependence model for the generalized propensity score incorrectly assumes conditional independence of exposures or is based on a naive dependence model which does not accommodate the effect of confounders on the conditional association of exposures. The cases of binary bivariate, continuous bivariate, and semi-continuous bivariate exposure are considered and a two-stage estimation technique is proposed to study the effects of prenatal alcohol exposure, and the effects of drinking frequency and intensity on childhood cognition. This chapter contains a summary of work appearing in Li et al (2024a).

Chapter 3 is motivated by studies of blood donation processes where the aim is to show that a new donation process does not lead to an increased risk of adverse events as introduced in Section 1.3.2 where the participants have blood donations repeatedly. Here a binary response is at the level of the transfusion and indicates whether an adverse event

was observed. Our interest is in study design (i.e. sample size calculations) as well as how to conduct interim analyses in order to assess the safety of the device on the response to transfusions. For blood donation the rate of serious hypotensive adverse events is very low so the log links are considered in the derivations of this chapter. The joint distribution of sequential test statistics for assessing the treatment effect over successive interim analyses is derived based on the method proposed by [Lee and Demets \(1991\)](#) for longitudinal data using generalized estimating equations and robust variance estimation. The number of repeated measurements is modelled using a zero-truncated negative binomial model with dependency on the recruitment time. We consider different alpha spending functions to determine the critical values of the test statistic for each interim analysis. Sample size formula is developed for the study design with planned interim analyses to ensure power requirements are met for the treatment effect on the binary response. We then conduct simulations to evaluate the empirical rejection rates for both the superiority trial design and the non-inferiority trial design. The research work of this chapter is summarized in [Li and Cook \(2024b\)](#) and [Li and Cook \(2024c\)](#).

In Chapter 4 we consider the design and analysis of two-phase studies aiming to assess the relationship between a biomarker and the disease progression under intermittent observations. We propose designs based on sub-sampling individuals with extreme score statistics which is initially introduced by [Tao et al \(2020\)](#) for continuous, discrete, or censored outcomes. Likelihood and estimating function approaches are developed and the expectation-maximization algorithm is used for the estimation of parameters. We conduct simulation studies to investigate the robustness and efficiency of the score-based residual sampling strategies for joint models of the marker and disease progression processes. An application is given to the study of patients with psoriatic arthritis as introduced in Section [1.4.2](#). The aim is to select individuals for the expensive measurements of the genetic marker HLA-B27 with the proposed selection strategy and evaluate the effect of this marker on the development of damaged joints which are measured at each clinical visit. This chapter contains a summary of work that has been prepared as a manuscript in [Li and Cook \(2024d\)](#).

Chapter 2

Generalized Propensity Scores for Assessing Effects of Multiple Exposures

2.1 Introduction

Causal analysis is concerned with mitigating the effect of confounding variables on inferences regarding the exposure-disease relationship. With binary exposures, propensity scores capturing the association between confounders and the exposure of interest can be used for matching, stratification, or regression adjustment ([Rosenbaum and Rubin, 1983](#)). Inverse probability of exposure weighting offers another framework for the estimation of causal effects. For continuous exposure variables, generalized propensity scores have been developed ([Imbens, 2000](#); [Hirano and Imbens, 2004](#)).

In settings with a single toxin, interest may lie in the causal effect of different exposure patterns defined in terms of the frequency of exposure, the dose of exposure on exposure occasions, and their interaction. [Di Credico et al \(2019\)](#) for example, study the effect of the duration and intensity of smoking on the risk of head and neck cancers. This is the setting of our motivating study described in Section [1.2.2](#) assessing the effect of both maternal drinking frequency and drinking intensity on childhood cognition.

When considering confounding in the causal analysis of multiple exposures, the effect of confounders is usually modeled on each marginal exposure variable with dependencies accommodated between the exposures through joint models - conditional associations are

typically assumed to be functionally independent of any confounders. We consider a more general framework in which confounders may affect second moments (i.e., the relation between different features of exposures to a single toxin). Specifically, we develop methods to study the joint effect of the frequency of drinking and the dose of alcohol per drinking occasion in expectant mothers on the cognitive development of their child (Jacobson et al, 1993); we then apply the methods we develop to the motivating study.

The remainder of Chapter 2 is as follows. In Section 2.2, we define a response model involving the main effects of two binary exposures and their interaction, along with a generalized propensity score. We formulate a general model for the exposures in which confounders affect the first and second-order moments so that they explain how the two aspects of exposure vary together. We then describe how this can be used for regression adjustment and inverse weighting. The importance of modeling the effect of confounders on the association between exposures is investigated in Section 2.3 by evaluating the bias of causal estimands from regression adjustment and inverse probability of exposure weighting when naive propensity scores only deal with the marginal effects of confounders. In Section 2.4, we consider the case of continuous exposures including semi-continuous exposures, and carry out an analogous investigation for analyses based on regression adjustment and inverse density weighting for estimating causal effects. A two-stage approach for estimating the causal effects of semi-continuous exposures is proposed in Section 2.5. An application to the Detroit Cohort Study (Jacobson et al, 2004) is given in Section 2.6 and concluding remarks are given in Section 2.7. We focus on the case of a bivariate exposure but as suggested by the notation in Section 2.2.1, the methods can be naturally generalized to accommodate higher dimensional exposures.

2.2 Causal Analysis with Multiple Exposures

2.2.1 Specification of the Full Model

Let Y denote a continuous response variable of interest, $\mathbf{A} = (A_1, \dots, A_p)'$ a $p \times 1$ vector of exposure variables with $p \geq 2$, and $\mathbf{B} = (A_1 A_2, \dots, A_{p-1} A_p)'$ a $q \times 1$ vector of all pairwise products where $q = p(p-1)/2$. Higher order interaction terms can be considered, but here we focus on two-way interactions and so let $\mathbf{X} = (\mathbf{A}', \mathbf{B}')$ contain the exposure data of interest. Furthermore, let $\mathbf{Z} = (Z_1, \dots, Z_r)'$ be an $r \times 1$ vector of confounding variables obscuring the effects of $\mathbf{X}$ on the response Y . The data generating model for the response is taken to have the form

$$Y = \mathbf{X}'\boldsymbol{\theta} + \mathbf{Z}'\boldsymbol{\zeta} + \epsilon, \quad (2.1)$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ with $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \dots, \alpha_p)'$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)'$, $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_r)'$, and ϵ is a mean zero error term with $\mathbf{X}, \mathbf{Z} \perp \epsilon$. Let $\mathbf{A}_{(-j)} = (A_1, \dots, A_{j-1}, A_{j+1}, \dots, A_p)'$ be the vector of all exposure variables excluding A_j , let $\mathbf{B}_{(-j)}$ be the vector of all products of A_j with the $p-1$ other exposures, and let $\boldsymbol{\beta}_{(-j)}$ be the corresponding vector of coefficients. The causal effect of a one unit change in A_j from a_{j0} to $a_{j0} + 1$ given $\mathbf{A}_{(-j)} = \mathbf{a}_{(-j)}$ and $\mathbf{Z}$ is $\alpha_j + \mathbf{a}_{(-j)}' \boldsymbol{\beta}_{(-j)}$. This model therefore enables one to study the causal interacting effects of the two exposures.

When dimension of $\mathbf{Z}$ is large or one is concerned about getting the functional form of the effects of confounders correct in the response model, regression adjustment on the propensity score mitigates the effects of confounders. Let $f_{a|z}(\mathbf{a} | \mathbf{z})$ be the conditional joint probability distribution or mass function for multivariate continuous or discrete exposures given confounders. Imai and Van Dyk (2004) defined the generalized propensity score (GPS) as $G(\mathbf{A}, \mathbf{Z}) = f_{a|z}(\mathbf{A} | \mathbf{Z})$ so that $\mathbf{Z} \perp I(\mathbf{A} = \mathbf{a}) | G(\mathbf{A}, \mathbf{Z}) = g$. Additionally, they stated that if there is a finite dimensional quantity $\mathbf{S}(\mathbf{Z})$ such that $G(\mathbf{A}, \mathbf{Z})$ depends only on $\mathbf{Z}$ through $\mathbf{S}(\mathbf{Z})$, matching or stratification on $\mathbf{S}(\mathbf{Z}) = \mathbf{s}$ or $G(\mathbf{A}, \mathbf{Z})$ is equivalent. For a single Bernoulli exposure variable, the generalized propensity score is the conventional propensity score, namely the conditional probability of being exposed given confounders (i.e., $P(A_1 = 1 | \mathbf{Z})$), whereas for a scalar continuous exposure it is the conditional density of the exposure $f_{a_1|z}(A_1 | \mathbf{Z})$; based on Assumption 3 of Imai and Van Dyk (2004) it will often be sufficient to condition on a linear function of confounders such as $\mathbb{E}(A_1 | \mathbf{Z})$. For multivariate exposures, the generalized propensity score is the multivariate joint distribution of exposures given confounders, $G(\mathbf{A}, \mathbf{Z}) = f_{a|z}(\mathbf{A} | \mathbf{Z})$. When the dependence of $\mathbf{A}$ on $\mathbf{Z}$ in $G(\mathbf{A}, \mathbf{Z})$ can be expressed through a low dimensional function $\mathbf{S}(\mathbf{Z})$, this generalized propensity score $\mathbf{S}(\mathbf{Z})$ can be considered the effective propensity score; we consider this in what follows.

2.2.2 Propensity Scores for Two Binary Exposures

Regression Adjustment

Here we aim to determine the minimal requirements for a potential vector-valued propensity score $\mathbf{S}(\mathbf{Z})$ to ensure valid causal inference regarding the exposure effects conditional on $\mathbf{S}(\mathbf{Z})$. We assume that data are available from independent individual processes and consider properties of a single process in much of what follows. From (2.1) note

$$\mathbb{E}_{Y|\mathbf{X}, \mathbf{S}}(Y) = \mathbb{E}_{\mathbf{Z}|\mathbf{X}, \mathbf{S}}\{\mathbb{E}_{Y|\mathbf{X}, \mathbf{S}, \mathbf{Z}}(Y)\} = \mathbf{X}'\boldsymbol{\theta} + \mathbb{E}_{\mathbf{Z}|\mathbf{X}, \mathbf{S}}(\mathbf{Z}')\boldsymbol{\zeta}. \quad (2.2)$$

Using $P(\cdot)$ to informally denote a joint probability mass function or density, we write

$$P(\mathbf{Z} \mid \mathbf{X}, \mathbf{S}) = \frac{P(\mathbf{X}, \mathbf{Z}, \mathbf{S})}{\int_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z}, \mathbf{S}) d\mathbf{z}} = \frac{P(\mathbf{X} \mid \mathbf{Z}, \mathbf{S})P(\mathbf{Z}, \mathbf{S})}{\int_{\mathbf{Z}} P(\mathbf{X} \mid \mathbf{Z}, \mathbf{S})P(\mathbf{Z}, \mathbf{S}) d\mathbf{z}}, \quad (2.3)$$

so if $\mathbf{S}(\mathbf{Z})$ contains enough information to construct the conditional distribution for the exposure variables given $\mathbf{Z}$ (i.e., if $\mathbf{X} \perp \mathbf{Z} \mid \mathbf{S}$), then following Assumption 3 of [Imai and Van Dyk \(2004\)](#), $\mathbf{S}$ serves as a valid propensity score. In this case the final term in (2.2) is independent of $\mathbf{X}$ and a working regression model of the form

$$\mathbb{E}(Y \mid \mathbf{X}, \mathbf{S}) = \mathbf{X}'\boldsymbol{\lambda} + \mathbf{S}'\boldsymbol{\eta} \quad (2.4)$$

will give an estimator $\hat{\boldsymbol{\lambda}}$ which is consistent for the causal parameter $\boldsymbol{\theta}$.

Here we give an illustrative example involving a bivariate binary exposure $\mathbf{A} = (A_1, A_2)'$ where A_k indicates exposure status with respect to exposure k , $k = 1, 2$. Let $\mathbf{Z}$ be a set of all potential confounders. Suppose the status for exposure k can be characterized by the marginal model $P(A_k \mid \mathbf{Z}) = P(A_k \mid \mathbf{Z}_k)$, where $\mathbf{Z}_k$ is a subvector of $\mathbf{Z}$, $k = 1, 2$. We may assume, for example, that $\text{logit}P(A_k = 1 \mid \mathbf{Z}_k) = \mathbf{Z}_k'\boldsymbol{\gamma}_k$. While such marginal models are typically used when considering scalar exposures, they may not be sufficient to fully characterize the joint distribution of exposures if some confounding variables exert a second-order effect. Let $\mathbf{Z}_3$ be another set of confounders characterizing the dependence between A_1 and A_2 given $\mathbf{Z}_1$ and $\mathbf{Z}_2$ through a dependence regression model

$$\log \text{OR}(A_1, A_2 \mid \mathbf{Z}) = \mathbf{Z}_3'\boldsymbol{\gamma}_3, \quad (2.5)$$

where

$$\text{OR}(A_1, A_2 \mid \mathbf{Z}) = \frac{P(A_1 = 1, A_2 = 1 \mid \mathbf{Z})/P(A_1 = 0, A_2 = 1 \mid \mathbf{Z})}{P(A_1 = 1, A_2 = 0 \mid \mathbf{Z})/P(A_1 = 0, A_2 = 0 \mid \mathbf{Z})}.$$

Based on the marginal models for $A_k \mid \mathbf{Z}_k$, $k = 1, 2$, and (2.5),

$$P(A_1 = 1, A_2 = 1 \mid \mathbf{Z}) = \begin{cases} \frac{a - [a^2 - 4\text{OR}(\text{OR} - 1)P(A_1=1|\mathbf{Z}_1)P(A_2=1|\mathbf{Z}_2)]^{1/2}}{2(\text{OR} - 1)}, & \text{OR} \neq 1 \\ P(A_1 = 1 \mid \mathbf{Z}_1)P(A_2 = 1 \mid \mathbf{Z}_2), & \text{OR} = 1 \end{cases}$$

where $a = 1 - (1 - \text{OR}) \sum_{k=1}^2 P(A_k = k \mid \mathbf{Z}_k)$ ([Lipsitz et al, 1991](#)) and we let OR denote $\text{OR}(A_1, A_2 \mid \mathbf{Z})$. The generalized propensity score is obtained by noting that $P(A_1 = 1, A_2 = 0 \mid \mathbf{Z}) = P(A_1 = 1 \mid \mathbf{Z}_1) - P(\mathbf{A} = \mathbf{1} \mid \mathbf{Z})$ and $P(A_1 = 0, A_2 = 1 \mid \mathbf{Z}) = P(A_2 = 1 \mid \mathbf{Z}_2) - P(\mathbf{A} = \mathbf{1} \mid \mathbf{Z})$, from which $P(\mathbf{A} = \mathbf{0} \mid \mathbf{Z})$ can be obtained. A sufficient propensity score would be $\mathbf{S}(\mathbf{Z}) = (S_1(\mathbf{Z}_1), S_2(\mathbf{Z}_2), S_3(\mathbf{Z}_3))'$ where $S_k(\mathbf{Z}_k) = P(A_k = 1 \mid \mathbf{Z}_k)$,

$k = 1, 2$ and $S_3(\mathbf{Z}) = P(\mathbf{A} = \mathbf{1} \mid \mathbf{Z})$; alternatively one could define $S_k(\mathbf{Z}_k) = \mathbf{Z}'_k \boldsymbol{\gamma}_k$, $k = 1, 2, 3$.

Here $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are sets of confounders influencing the marginal causal effects of A_1 and A_2 on Y , and conditioning on (S_1, S_2) is sufficient if A_1 does not interact with A_2 (i.e., if $\beta_1 = 0$). If $\beta_1 \neq 0$, then it is necessary to model the association between A_1 and A_2 given $(\mathbf{Z}'_1, \mathbf{Z}'_2)'$ as a function of possibly different sets of confounders $\mathbf{Z}_3$ as suggested in (2.5). We investigate the implications of not doing this adequately in Section 2.3.

Inverse Probability Weighting

Here we consider the use of a generalized propensity score as weight in inverse probability weighting analysis for two binary exposures. Let $P(\mathbf{A} \mid \mathbf{Z}; \boldsymbol{\gamma})$ denote the conditional joint probability of $\mathbf{A} = (A_1, A_2)'$ given confounders under the positivity condition (Rosenbaum and Rubin, 1983) where $P(\mathbf{A} = \mathbf{a} \mid \mathbf{Z}; \boldsymbol{\gamma}) > 0$ for all possible values of $\mathbf{Z}$. The inverse “joint probability” weighted estimating function (Hernán and Robins, 2010) is expressed as

$$U_1(\boldsymbol{\varphi}; \boldsymbol{\gamma}) = \frac{I(\mathbf{A} = \mathbf{a})}{P(\mathbf{A} = \mathbf{a} \mid \mathbf{Z}; \boldsymbol{\gamma})} D(\mathbf{X}) \{Y - \mathbb{E}^R(Y \mid \mathbf{X}; \boldsymbol{\varphi})\}, \quad (2.6)$$

where $\mathbb{E}^R(\cdot)$ represents an expectation in the setting where $\mathbf{X} \perp \mathbf{Z}$ and $\mathbb{E}^R(Y \mid \mathbf{X}; \boldsymbol{\varphi}) = \mathbf{X}'\boldsymbol{\varphi}$ defines a marginal structural model. Here

$$D(\mathbf{X}) = \frac{\partial}{\partial \boldsymbol{\varphi}} \mathbb{E}^R(Y \mid \mathbf{X}; \boldsymbol{\varphi}) = (1, A_1, A_2, A_1 A_2)'.$$

To show that (2.6) is an unbiased estimating function, we take the expectation of $U_1(\boldsymbol{\varphi}; \boldsymbol{\gamma})$ with respect to $Y \mid \mathbf{X}, \mathbf{Z}$ to obtain

$$\frac{I(\mathbf{A} = \mathbf{a})}{P(\mathbf{A} = \mathbf{a} \mid \mathbf{Z}; \boldsymbol{\gamma})} D(\mathbf{X}) \{\mathbb{E}(Y \mid \mathbf{X}, \mathbf{Z}; \Psi) - \mathbb{E}^R(Y \mid \mathbf{X}; \boldsymbol{\varphi})\}, \quad (2.7)$$

where $\Psi = (\boldsymbol{\theta}', \boldsymbol{\zeta}')'$. The expectation of (2.7) with respect to $\mathbf{Z} \mid \mathbf{A} = \mathbf{a}$ gives

$$\sum_{\mathbf{Z}} \frac{I(\mathbf{A} = \mathbf{a})}{P(\mathbf{A} = \mathbf{a} \mid \mathbf{Z}; \boldsymbol{\gamma})} D(\mathbf{X}) \{\mathbb{E}(Y \mid \mathbf{X}, \mathbf{Z}; \Psi) - \mathbb{E}^R(Y \mid \mathbf{X}; \boldsymbol{\varphi})\} P(\mathbf{Z} \mid \mathbf{A} = \mathbf{a}). \quad (2.8)$$

Since

$$P(\mathbf{Z} \mid \mathbf{A} = \mathbf{a}) = \frac{P(\mathbf{A} = \mathbf{a} \mid \mathbf{Z}; \boldsymbol{\gamma}) P(\mathbf{Z}; \boldsymbol{\tau})}{P(\mathbf{A} = \mathbf{a})},$$

we can rewrite (2.8) as

$$\sum_{\mathbf{Z}} \frac{I(\mathbf{A} = \mathbf{a})}{P(\mathbf{A} = \mathbf{a})} D(\mathbf{X}) \{ \mathbb{E}(Y | \mathbf{X}, \mathbf{Z}; \Psi) - \mathbb{E}^R(Y | \mathbf{X}; \boldsymbol{\varphi}) \} P(\mathbf{Z}; \boldsymbol{\tau}) ,$$

which effectively results in the expectation being taken in the setting where $\mathbf{Z} \perp \mathbf{A}$, equivalent to how this would be done in the hypothetical randomized trial (Hernán and Robins, 2010). Since (2.6) has expectation zero, solving $U_1(\boldsymbol{\varphi}; \boldsymbol{\gamma}) = 0$ defines an estimator that is consistent for the causal estimand of interest. In practice, $\boldsymbol{\gamma}$ must be estimated but provided a $\sqrt{n}$ -consistent estimator is used, consistency of $\hat{\boldsymbol{\varphi}}$ is still achieved.

2.3 Exposure Dependence Model Misspecification

2.3.1 Regression Adjustment

General Framework

Here we derive the limiting bias of the estimators of exposure effects obtained through regression adjustment when the dependence model is misspecified through a) an inappropriate conditional independence assumption, or b) specification of an exchangeable dependence structure functionally independent of confounders; this is defined more formally in what follows.

Consider the following regression model with a generalized propensity score adjustment,

$$\mathbb{E}(Y | \mathbf{X}, \tilde{\mathbf{S}}) = \mathbf{X}'\boldsymbol{\varphi} + \tilde{\mathbf{S}}'\boldsymbol{\psi} , \quad (2.9)$$

where $\tilde{\mathbf{S}} = \tilde{\mathbf{S}}(\mathbf{Z})$ is a vector-valued propensity score which might be insufficient for adjustment under model misspecification, e.g., assuming conditional independence among exposures or using an inadequate dependence model. The contribution to the least squares equation for $\boldsymbol{\varphi}$ from a single individual based on (2.9) is $U(Y | \mathbf{X}, \tilde{\mathbf{S}}; \boldsymbol{\varphi}, \boldsymbol{\psi}) = (\mathbf{X}', \tilde{\mathbf{S}}')(Y - \mathbf{X}'\boldsymbol{\varphi} - \tilde{\mathbf{S}}'\boldsymbol{\psi})$. With a sample of n independent individuals yielding data $\{(Y_i, \mathbf{X}_i, \mathbf{Z}_i), i = 1, \dots, n\}$, the least squares estimator of the exposure effect is the solution to $\sum_{i=1}^n U(Y_i | \mathbf{X}_i, \tilde{\mathbf{S}}_i; \boldsymbol{\varphi}, \boldsymbol{\psi}) = 0$ and is denoted by $\hat{\boldsymbol{\varphi}}$; this is consistent for the solution to $\mathbb{E}\{U(Y | \mathbf{X}, \tilde{\mathbf{S}}; \boldsymbol{\varphi}, \boldsymbol{\psi})\} = 0$ where the expectation is taken with respect to the full data generation process (White, 1982; Rotnitzky and Wypij, 1994).

Again we consider the bivariate case with $\mathbf{X} = (1, A_1, A_2, A_1A_2)'$ and the corresponding coefficient $\boldsymbol{\varphi} = (\varphi_0, \varphi_1, \varphi_2, \varphi_3)'$. The explicit form of the limiting value of the estimator $\tilde{\boldsymbol{\varphi}}$ for $\boldsymbol{\theta}$ in (2.1) is

$$\tilde{\varphi}_1 = \alpha_1 + \boldsymbol{\zeta}'\Delta_{11} + \boldsymbol{\psi}'\Delta_{21}, \quad (2.10)$$

$$\tilde{\varphi}_2 = \alpha_2 + \boldsymbol{\zeta}'\Delta_{12} + \boldsymbol{\psi}'\Delta_{22}, \quad (2.11)$$

$$\tilde{\varphi}_3 = \beta_1 + \boldsymbol{\zeta}'\left(\frac{\Sigma_3}{\sigma_3^2} - \Delta_{11}\frac{\rho_{13}\sigma_1}{\sigma_3} - \Delta_{12}\frac{\rho_{23}\sigma_2}{\sigma_3}\right) - \boldsymbol{\psi}'\left(\frac{\Lambda_3}{\sigma_3^2} + \Delta_{21}\frac{\rho_{13}\sigma_1}{\sigma_3} + \Delta_{22}\frac{\rho_{23}\sigma_2}{\sigma_3}\right), \quad (2.12)$$

where $\boldsymbol{\zeta}$ and $\boldsymbol{\psi}$ are given in (2.1) and (2.9),

$$\Delta_{1k} = \frac{(\rho_{12} - \rho_{13}\rho_{23})\left(\frac{\sigma_k^2\Sigma_{3-k}}{\sigma_1\sigma_2} - \frac{\sigma_k}{\sigma_3}\frac{\rho_{13}\rho_{23}}{\rho_{k3}}\Sigma_3\right) + (\frac{\rho_{13}^2\rho_{23}^2}{\rho_{k3}^2} - 1)(\Sigma_k - \frac{\sigma_k}{\sigma_3}\rho_{k3}\Sigma_3)}{\sigma_k^2\{(\rho_{12} - \rho_{13}\rho_{23})^2 - (1 - \rho_{13}^2)(1 - \rho_{23}^2)\}},$$

$$\Delta_{2k} = \frac{(\rho_{12} - \rho_{13}\rho_{23})\left(\frac{\sigma_k^2\Lambda_{3-k}}{\sigma_1\sigma_2} - \frac{\sigma_k}{\sigma_3}\frac{\rho_{13}\rho_{23}}{\rho_{k3}}\Lambda_3\right) + (\frac{\rho_{13}^2\rho_{23}^2}{\rho_{k3}^2} - 1)(\Lambda_k - \frac{\sigma_k}{\sigma_3}\rho_{k3}\Lambda_3)}{\sigma_k^2\{(\rho_{12} - \rho_{13}\rho_{23})^2 - (1 - \rho_{13}^2)(1 - \rho_{23}^2)\}},$$

$\Sigma_k = \text{cov}(\mathbb{E}(A_k | \mathbf{Z}), \mathbf{Z})$, $\Lambda_k = \text{cov}(\mathbb{E}(A_k | \mathbf{Z}), \tilde{\mathbf{S}})$, $\Sigma_3 = \text{cov}(\mathbb{E}(A_1A_2 | \mathbf{Z}), \mathbf{Z})$ and $\Lambda_3 = \text{cov}(\mathbb{E}(A_1A_2 | \mathbf{Z}), \tilde{\mathbf{S}})$. We also have $\sigma_k^2 = \text{var}(A_k)$, $\sigma_3^2 = \text{var}(A_1A_2)$, $\rho_{12} = \text{corr}(A_1, A_2)$, and $\rho_{k3} = \text{corr}(A_k, A_1A_2)$, $k = 1, 2$; see Appendix A.1 for more details.

With two binary exposures, $\sigma_k^2 = p_k(1 - p_k)$ and $\sigma_3^2 = p_{12}(1 - p_{12})$, where $p_k = P(A_k = 1)$, $k = 1, 2$ and $p_{12} = P(A_1A_2 = 1)$. The correlation ρ_{12} is

$$\rho_{12} = \frac{p_{12} - p_1p_2}{\sqrt{p_1(1 - p_1)p_2(1 - p_2)}}$$

(Yule, 1912), and

$$\rho_{k3} = \frac{p_{12} - p_kp_{12}}{\sqrt{p_k(1 - p_k)p_{12}(1 - p_{12})}}, \quad k = 1, 2.$$

A Misspecified Dependence Model

If exposure variables are treated as conditionally independent given confounders, one might specify $\tilde{\mathbf{S}}$ as $\tilde{\mathbf{S}} = (S_1(\mathbf{Z}_1), S_2(\mathbf{Z}_2))'$ or $\tilde{\mathbf{S}} = (S_1(\mathbf{Z}_1), S_2(\mathbf{Z}_2), S_1(\mathbf{Z}_1)S_2(\mathbf{Z}_2))'$, where $S_1(\mathbf{Z}_1)$ and $S_2(\mathbf{Z}_2)$ are the marginal propensity scores. A naive model using regression adjustment might be specified as

$$\mathbb{E}(Y | \mathbf{X}, \mathbf{S}) = \mathbf{X}'\boldsymbol{\varphi} + \psi_1S_1(\mathbf{Z}_1) + \psi_2S_2(\mathbf{Z}_2), \quad (2.13)$$

or possibly

$$\mathbb{E}(Y \mid \mathbf{X}, \mathbf{S}) = \mathbf{X}'\boldsymbol{\varphi} + \psi_1 S_1(\mathbf{Z}_1) + \psi_2 S_2(\mathbf{Z}_2) + \psi_3 S_1(\mathbf{Z}_1) S_2(\mathbf{Z}_2) . \quad (2.14)$$

In the most general setting such propensity score regression adjustments are insufficient to render $\mathbf{X} \perp \mathbf{Z} \mid \tilde{\mathbf{S}}$ since $A_1 A_2$ depends on the joint model of A_1 and A_2 given $\mathbf{Z}$. As a result, estimates of the causal effects φ_1, φ_2 and φ_3 are inconsistent with the asymptotic bias given by the second and third terms on the right-hand side of equations (2.10) to (2.12). It can be seen that the bias depends on the second-order association between the exposures and confounders, the association between exposures and generalized propensity scores, and variances of exposures.

A more flexible exposure model may recognize the conditional association between exposures (given marginal confounders $\mathbf{Z}_1$ and $\mathbf{Z}_2$), but assume no role of confounders on the second-order dependence. In this case, one might consider a propensity score term based on

$$\mathbb{E}(A_1 A_2 \mid \mathbf{Z}) = \text{cov}(A_1, A_2 \mid \mathbf{Z}_1, \mathbf{Z}_2) + S_1(\mathbf{Z}_1) S_2(\mathbf{Z}_2) ,$$

where $\text{cov}(A_1, A_2 \mid \mathbf{Z}_1, \mathbf{Z}_2) = \rho \sqrt{V(A_1 \mid \mathbf{Z}_1) V(A_2 \mid \mathbf{Z}_2)}$ and $V(A_k \mid \mathbf{Z}_k)$ is the residual variance from the propensity score for exposure k , $k = 1, 2$. In this case, the regression model adjusted by the generalized propensity score may be,

$$\mathbb{E}(Y \mid \mathbf{X}, \mathbf{S}) = \mathbf{X}'\boldsymbol{\varphi} + \psi_1 S_1(\mathbf{Z}_1) + \psi_2 S_2(\mathbf{Z}_2) + \psi_3 \mathbb{E}(A_1 A_2 \mid \mathbf{Z}_1, \mathbf{Z}_2) . \quad (2.15)$$

We explore the impact of such misspecifications in Section 2.3.3.

2.3.2 Inverse Probability Weighting

A misspecified propensity score will have a different impact in the framework of the inverse propensity score weighted estimating function. Let $P_0(\mathbf{A} = \mathbf{a} \mid \mathbf{Z}; \delta)$ denote the misspecified (naive) conditional joint density or mass function indexed by δ , and let $U_{10}(\boldsymbol{\varphi}; \delta)$ and $U_{20}(\delta)$ be the estimating functions as follows,

$$U_{10}(\boldsymbol{\varphi}; \delta) = \frac{I(\mathbf{A} = \mathbf{a})}{P_0(\mathbf{A} = \mathbf{a} \mid \mathbf{Z}; \delta)} D(\mathbf{X}) \{Y - \mathbb{E}^R(Y \mid \mathbf{X}; \boldsymbol{\varphi})\} , \quad (2.16)$$

$$U_{20}(\delta) = \frac{\partial \log P_0(\mathbf{A} \mid \mathbf{Z}; \delta)}{\partial \delta} . \quad (2.17)$$

Let $\boldsymbol{\delta}^*$ be the solution to $\mathbb{E}_{\mathbf{A}, \mathbf{Z}}\{U_{20}(\boldsymbol{\delta}); \boldsymbol{\gamma}, \boldsymbol{\tau}\} = 0$, the limiting value to which the estimator $\hat{\boldsymbol{\delta}}$ converges when fitting the misspecified model. The limiting value of $\boldsymbol{\varphi}$ is obtained by setting the equation below to zero and solving $\hat{\boldsymbol{\varphi}}$,

$$\sum_{\mathbf{A}} \sum_{\mathbf{Z}} \frac{I(\mathbf{A} = \mathbf{a}) P(\mathbf{A} = \mathbf{a} \mid \mathbf{Z}; \boldsymbol{\gamma})}{P_0(\mathbf{A} = \mathbf{a} \mid \mathbf{Z}; \boldsymbol{\delta}^*) P(\mathbf{Z}; \boldsymbol{\tau})} D(\mathbf{X}) \{ \mathbb{E}(Y \mid \mathbf{X}, \mathbf{Z}; \Psi) - \mathbb{E}^R(Y \mid \mathbf{X}; \boldsymbol{\varphi}) \}. \quad (2.18)$$

A misspecified propensity score $P_0(\mathbf{A} = \mathbf{a} \mid \mathbf{Z}; \boldsymbol{\delta})$ is the product of two marginal propensity scores $P(A_1 = a_1 \mid \mathbf{Z}; \boldsymbol{\delta})P(A_2 = a_2 \mid \mathbf{Z}; \boldsymbol{\delta})$ under the conditional independence assumption. In this case under the present formulation, we have

$$U_{20}(\boldsymbol{\delta}) = (U'_{20,1}(\boldsymbol{\gamma}_1), U'_{20,2}(\boldsymbol{\gamma}_2)) ,$$

where

$$U_{20,k}(\boldsymbol{\gamma}_k) = \frac{\partial \log P(A_k = a_k \mid \mathbf{Z}_k; \boldsymbol{\gamma}_k)}{\partial \boldsymbol{\gamma}_k} , \quad k = 1, 2 ,$$

and $\boldsymbol{\delta}^* = (\boldsymbol{\gamma}'_1, \boldsymbol{\gamma}'_2)'$ so $\hat{\boldsymbol{\delta}}$ is consistent for the target parameters $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ but is of insufficient dimension to characterize the second-order dependence.

When the association is assumed functionally independent of confounders, the misspecified propensity score $P_0(\mathbf{A} = \mathbf{a} \mid \mathbf{Z}; \boldsymbol{\delta})$ is an inadequately modeled conditional joint density or mass function $P(A_1 = a_1, A_2 = a_2 \mid \mathbf{Z}_1, \mathbf{Z}_2; \boldsymbol{\delta})$. In this case we have $U_{20}(\boldsymbol{\delta}) = \partial \log P(\mathbf{A} = \mathbf{a} \mid \mathbf{Z}_1, \mathbf{Z}_2; \boldsymbol{\delta}) / \partial \boldsymbol{\delta}$, and $\boldsymbol{\delta}^*$ will not be consistent for the target parameters $\boldsymbol{\gamma}_1$ or $\boldsymbol{\gamma}_2$.

2.3.3 Empirical Bias and Efficiency

Here we report on simulation studies investigating the limiting values of the exposure effects subject to model misspecification. We consider a bivariate binary exposure and assess the relative asymptotic bias of estimates of exposure effects under two types of model misspecification of Section 2.3.1 under both regression adjustment and IPW analyses.

Suppose a sample of n independent replicates $\{(Y_i, \mathbf{X}_i, \mathbf{Z}_i), i = 1 \dots n\}$ are generated with $n = 500$. Let $\mathbf{Z}$ contain binary confounding variables with $Z_1 \perp Z_2$, $P(Z_1 = 1) = P(Z_2 = 1) = 0.5$, and $P(Z_3 = 1 \mid Z_1, Z_2) = \text{expit}\{-0.896 + (\log 2)Z_1 + (\log 3)Z_2\}$ giving $P(Z_3 = 1) = 0.5$. We let $P(A_1 = 1 \mid Z_1) = \text{expit}(\gamma_{10} + \gamma_{11}Z_1)$ where $\boldsymbol{\gamma}_1 = (\gamma_{10}, \gamma_{11})' = (-0.764, \log 2)'$ so $\mathbb{E}(A_1) = 0.4$, and $P(A_2 = 1 \mid Z_2) = \text{expit}(\gamma_{20} + \gamma_{21}Z_2)$ with $\boldsymbol{\gamma}_2 = (\gamma_{20}, \gamma_{21})' = (-0.112, \log 1.25)'$ so $\mathbb{E}(A_2) = 0.5$. We let $\log \text{OR}(A_1, A_2 \mid Z_3) = \gamma_{30} + \gamma_{31}Z_3$

with $\boldsymbol{\gamma}_3 = (\gamma_{30}, \gamma_{31})'$; we consider $\gamma_{31} \in [-5, 5]$ and for each value of γ_{31} we solve for γ_{30} in (2.5) such that $\mathbb{E}\{\text{OR}(A_1, A_2)\} = 4$. For response Y we generate data according to (2.1) with $\boldsymbol{\theta} = (\alpha_0, \alpha_1, \alpha_2, \beta_1)' = (0, 2, 2, 4)'$, and $\boldsymbol{\zeta} = (0.5, 0.5, 0.5)'$. The data generation process is repeated for $n_{sim} = 1000$ times for the analyses as follow.

Under the assumption of conditional independence between exposures, the regression model adjusted by the generalized propensity scores can be fitted as in (2.13) or (2.14), which is denoted as Reg1 and Reg2 in the Figure 2.1 respectively. The IPW approach involves solving the exposure effect in (2.18). We denote the estimated effect as $\hat{\boldsymbol{\varphi}} = (\hat{\varphi}_1, \hat{\varphi}_2, \hat{\varphi}_3)'$ for $\mathbf{X} = (A_1, A_2, A_1A_2)'$, and the relative bias of the estimated exposure effect for $\mathbf{X}$ as $(\hat{\varphi}_1 - \alpha_1)/\alpha_1$, $(\hat{\varphi}_2 - \alpha_2)/\alpha_2$, and $(\hat{\varphi}_3 - \beta_1)/\beta_1$. The average relative bias is shown in Figure 2.1. In the bottom right panel of Figure 2.1, we show the relative bias of the causal effect of a one-unit change in A_2 given $A_1 = a_1 = 1$ or 0 and $\mathbf{Z}$, and denote it as $(\hat{\Delta}_2(A_1) - \Delta_2(A_1))/\Delta_2(A_1)$ where $\Delta_2(A_1) = \alpha_2 + a_1\beta_1$ and $\hat{\Delta}_2(A_1) = \hat{\varphi}_2 + a_1\hat{\varphi}_3$. The absolute value of relative bias of the exposure effect increases as the covariate effect on the exposure dependence increases. Moreover the regression adjustment leads to less bias in effect estimation than the IPW approach in these settings. Adjusting for $(S_1(\mathbf{Z}), S_2(\mathbf{Z}))'$ or $(S_1(\mathbf{Z}), S_2(\mathbf{Z}), S_1(\mathbf{Z})S_2(\mathbf{Z}))'$ produce similar results. When $\gamma_{31} = 0$ (i.e., the conditional association between exposure is functionally independent of confounders), the bias of the estimated exposure effect is negligible when conditional independence is assumed between two binary exposures.

We follow (2.15) and (2.18) to conduct the estimation for the second type of misspecification, that is using an exchangeable dependence structure which assumes no second-order effects of confounders on the exposure. The results are shown in Figure 2.2. When there is no confounder affecting the dependence (i.e., $\gamma_{31} = 0$), the curves pass through zero points on both axes, indicating both the regression adjustment and IPW approaches produce unbiased estimates as the models are correctly specified under the assumption of no second-order effects of confounders on the exposure. Similarly as in Figure 2.1, the absolute value of relative bias of the exposure effect would increase as the covariate effect on the exposure dependence (γ_{31}) increases. The regression adjustment leads to less bias in effect estimation compared to the IPW approach.

2.4 Multivariate Continuous Exposures

In this section, we consider multiple continuous exposures and semi-continuous exposures.

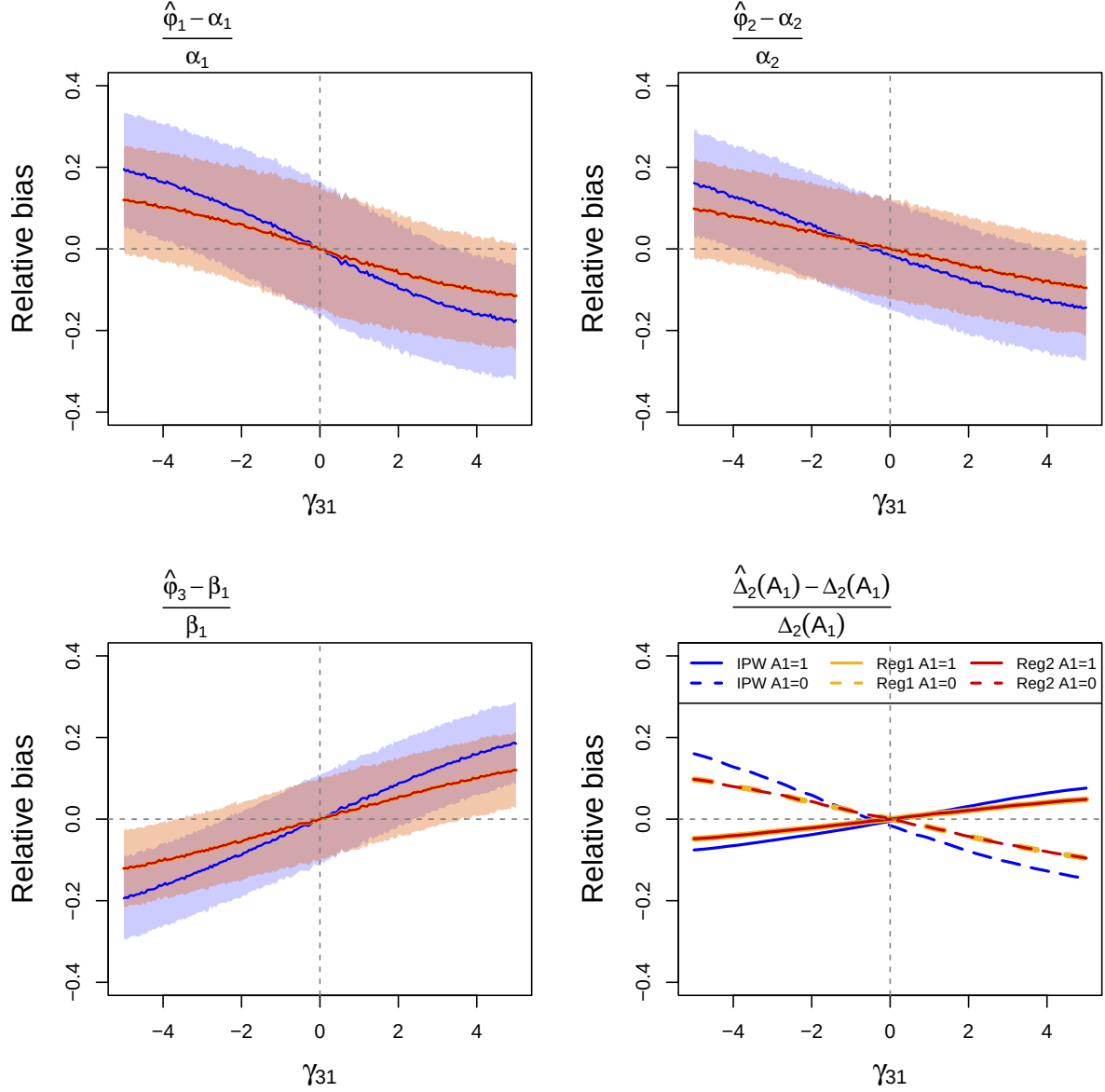


Figure 2.1: Empirical relative bias with standard errors (shaded bands) of estimates of exposure effects from inadequate propensity score modeling assuming conditional independence between two binary exposures when the second-order confounder has the effect γ_{31} ; regression adjustment (Reg) (yellow+red) and inverse weighting (IPW) (blue) are considered; the marginal association between two binary exposures is $\mathbb{E}\{\text{OR}(A_1, A_2)\} = 4$.

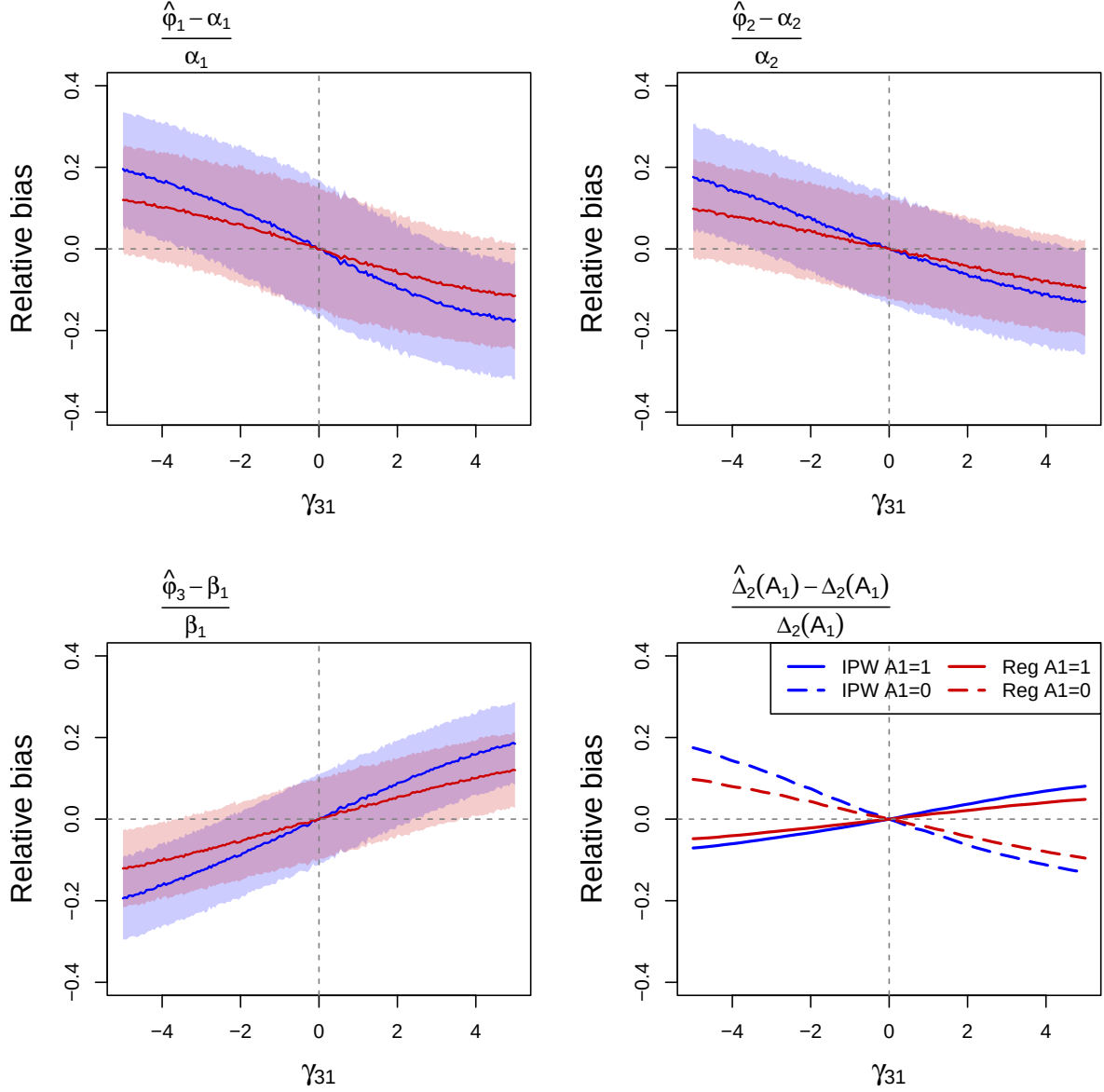


Figure 2.2: Empirical relative bias with standard errors (shaded bands) of estimates of exposure effects from inadequate propensity score modeling due to the omission of a second-order confounder with effect γ_{31} ; regression adjustment (Reg) (red) and inverse weighting (IPW) (blue) are considered; the marginal association between two binary exposures is $\mathbb{E}\{\text{OR}(A_1, A_2)\} = 4$.

2.4.1 Multivariate Normal Exposures

Regression Adjustment

Here we consider $\mathbf{A} \mid \mathbf{Z} \sim \text{BVN}(\boldsymbol{\mu}(\mathbf{Z}), \Sigma(\mathbf{Z}))$ with $\mathbf{A} = (A_1, A_2)'$, $\mathbf{Z} = (\mathbf{Z}'_1, \mathbf{Z}'_2, \mathbf{Z}'_3)'$, $\boldsymbol{\mu}(\mathbf{Z}) = (\mu_1(\mathbf{Z}), \mu_2(\mathbf{Z}))'$, $\mu_k(\mathbf{Z}) = \mathbb{E}(A_k \mid \mathbf{Z}) = \mathbb{E}(A_k \mid \mathbf{Z}_k)$, $\text{var}(A_k \mid \mathbf{Z}) = \sigma_k^2$, $k = 1, 2$, and $\text{corr}(A_1, A_2 \mid \mathbf{Z}) = \rho(\mathbf{Z}_3)$. We suppose $\mathbb{E}(A_k \mid \mathbf{Z}) = \mathbf{Z}'_k \boldsymbol{\gamma}_k$ with a dependence model

$$\log \left\{ \frac{1 + \rho(\mathbf{Z}_3)}{1 - \rho(\mathbf{Z}_3)} \right\} = \mathbf{Z}'_3 \boldsymbol{\gamma}_3. \quad (2.19)$$

Since the variance parameters are functionally independent of $\mathbf{Z}$, a sufficient propensity score is $\mathbf{S}(\mathbf{Z}) = (S_1(\mathbf{Z}_1), S_2(\mathbf{Z}_2), S_3(\mathbf{Z}_3))'$ where $S_k(\mathbf{Z}_k) = \mathbf{Z}'_k \boldsymbol{\gamma}_k$, $k = 1, 2$ and $S_3(\mathbf{Z}) = \mathbf{Z}'_3 \boldsymbol{\gamma}_3$ (Imai and Van Dyk, 2004) since

$$f(a_1, a_2 \mid \mathbf{Z}, \mathbf{S}) = \frac{1}{2\pi \sqrt{|\Sigma(\mathbf{S}_3)|}} \exp\{-(\mathbf{a} - \mathbf{S})' \Sigma^{-1}(\mathbf{S}_3)(\mathbf{a} - \mathbf{S})/2\},$$

where $(\mathbf{a} - \mathbf{S}) = (a_1 - S_1, a_2 - S_2)'$ and

$$\Sigma(\mathbf{S}_3) = \begin{bmatrix} \sigma_1^2 & \rho(\mathbf{Z}_3) \sigma_1 \sigma_2 \\ \rho(\mathbf{Z}_3) \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix}.$$

An alternative to (2.19) would be to model $B_1 = A_1 A_2$ given $\mathbf{Z}$ directly to define $S_3(\mathbf{Z}) = \mathbb{E}(B_1 \mid \mathbf{Z}_3)$ and note that in this case we could replace the covariance entries of Σ with $S_3(\mathbf{Z}) - S_1(\mathbf{Z}_1)S_2(\mathbf{Z}_2)$; the key point is that $\mathbf{S}(\mathbf{Z}) = (S_1(\mathbf{Z}_1), S_2(\mathbf{Z}_2), S_3(\mathbf{Z}_3))'$ contains the relevant information on confounders in the exposure model.

Inverse Joint Density Weighting

Inverse density weighting can be used when exposures are continuous (Williams and Crespi, 2020). With conditional joint density $f(\mathbf{a} \mid \mathbf{z}; \boldsymbol{\gamma})$, as in (2.6), we write the inverse joint density weighted estimating function as

$$U_1(\boldsymbol{\varphi}; \boldsymbol{\gamma}) = \frac{I(\mathbf{A} = \mathbf{a})}{f(\mathbf{a} \mid \mathbf{z}; \boldsymbol{\gamma})} D(\mathbf{X}) \{Y - \mathbb{E}^R(Y \mid \mathbf{X}; \boldsymbol{\varphi})\}. \quad (2.20)$$

The expectation of $U_1(\boldsymbol{\varphi}; \boldsymbol{\gamma})$ with respect to $Y \mid \mathbf{X}, \mathbf{Z}$ and $\mathbf{Z} \mid \mathbf{X}$ gives

$$\int_{\mathbf{Z}} \frac{I(\mathbf{A} = \mathbf{a})}{f(\mathbf{a} \mid \mathbf{z}; \boldsymbol{\gamma})} D(\mathbf{X}) \{ \mathbb{E}(Y \mid \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}, \boldsymbol{\zeta}) - \mathbb{E}^R(Y \mid \mathbf{X}; \boldsymbol{\varphi}) \} f(\mathbf{z} \mid \mathbf{a}) d\mathbf{z} , \quad (2.21)$$

where

$$f(\mathbf{z} \mid \mathbf{a}) = \frac{f(\mathbf{a} \mid \mathbf{z}; \boldsymbol{\gamma}) f(\mathbf{z}; \boldsymbol{\tau})}{f(\mathbf{a})} .$$

Thus, we can rewrite (2.21) as

$$\int_{\mathbf{Z}} \frac{I(\mathbf{A} = \mathbf{a})}{f(\mathbf{a})} D(\mathbf{X}) \{ \mathbb{E}(Y \mid \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}, \boldsymbol{\zeta}) - \mathbb{E}^R(Y \mid \mathbf{X}; \boldsymbol{\varphi}) \} f(\mathbf{z}; \boldsymbol{\tau}) d\mathbf{z} , \quad (2.22)$$

which results in the expectation being taken in the setting where $\mathbf{Z} \perp \mathbf{A}$, and gives consistent causal estimate $\hat{\boldsymbol{\varphi}}$ from solving $U_1(\boldsymbol{\varphi}; \boldsymbol{\gamma}) = 0$.

2.4.2 Multivariate Semi-continuous Exposures

Regression Adjustment

Here we suppose A_1 and A_2 are non-negative semi-continuous exposure variables with a mass at zero and we adopt a two-part generalized propensity score model (Akkaya Hocagil et al, 2021) to address the effects of confounding variables. While these two exposures could in general represent exposures to different agents, here we consider the case where they represent different aspects of the exposure to a single agent. Specifically for motivating study we may let A_1 and A_2 represent the drinking frequency (proportion of drinking days per week) and drinking intensity (dose per drinking occasion) – among non-drinkers these degenerate to zero. We can therefore let $A^+ = \mathbb{I}(A_1 A_2 > 0)$ indicate positive exposure, $L_k = \log(A_k \mid A^+ = 1)$, $k = 1, 2$, and $\pi_1(\mathbf{Z}_0) = \mathbb{E}(A^+ \mid \mathbf{Z}_0)$.

We consider the true response model is (2.1) with $\mathbf{X} = (1, A^+, A^+ L_1, A^+ L_2, A^+ L_1 L_2)'$. Let

$$\pi_1(\mathbf{Z}_0) = \text{expit}(\mathbf{Z}_0' \boldsymbol{\gamma}_0) ,$$

where $\mathbf{Z}_0 = (1, \mathbf{Z}_1', \mathbf{Z}_2')'$, and given $A^+ = 1$, we suppose $L_k = \mathbf{Z}_k' \boldsymbol{\gamma}_k + R_k$, $k = 1, 2$ and $\mathbf{R} = (R_1, R_2)'$ is bivariate normal with $\text{var}(R_k) = \sigma_k^2$, $k = 1, 2$, $\rho(\mathbf{Z}) = \text{corr}(R_1, R_2 \mid \mathbf{Z})$

and $\log\{(1 + \rho(\mathbf{Z})/(1 - \rho(\mathbf{Z}))\} = \mathbf{Z}'_3\boldsymbol{\gamma}_3$. As such, we have

$$\begin{aligned}\mathbb{E}(L_k \mid \mathbf{Z}_0, \mathbf{Z}_k) &= \pi_1(\mathbf{Z}_0)\mathbb{E}(L_k \mid A^+ = 1, \mathbf{Z}_k), \quad k = 1, 2, \\ \mathbb{E}(L_1 L_2 \mid \mathbf{Z}) &= \pi_1(\mathbf{Z}_0)\mathbb{E}(L_1 L_2 \mid A^+ = 1, \mathbf{Z})\end{aligned}$$

to form the generalized propensity score as $\mathbf{S} = (\pi_1(\mathbf{Z}_0), \mathbb{E}(L_1 \mid \mathbf{Z}_0, \mathbf{Z}_1), \mathbb{E}(L_2 \mid \mathbf{Z}_0, \mathbf{Z}_2), \mathbb{E}(L_1 L_2 \mid \mathbf{Z}))'$ or represented as $\mathbf{S} = (S_0(\mathbf{Z}_0), S_1(\mathbf{Z}_1), S_2(\mathbf{Z}_2), S_3(\mathbf{Z}_3))'$ and use it for adjustment in the regression model (2.4) as

$$\mathbb{E}(Y \mid \mathbf{X}, \mathbf{S}) = \mathbf{X}'\boldsymbol{\lambda} + \mathbf{S}'\boldsymbol{\eta}, \quad (2.23)$$

where $\mathbf{X} = (1, A^+, A^+L_1, A^+L_2, A^+L_1L_2)'$.

Inverse Joint Density Weighting

The inverse density of exposure weighted estimating function is given here by

$$U_0(Y \mid \mathbf{X}; \boldsymbol{\varphi}, \boldsymbol{\Gamma}, \boldsymbol{\gamma}_0) = w(\mathbf{X}, \mathbf{Z})D(\mathbf{X}) \{Y - \mathbb{E}^R(Y \mid \mathbf{X}; \boldsymbol{\varphi})\}, \quad (2.24)$$

where $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}'_1, \sigma_1^2, \boldsymbol{\gamma}'_2, \sigma_2^2, \boldsymbol{\gamma}'_3)'$,

$$w(\mathbf{X}, \mathbf{Z}) = \frac{\mathbb{I}(\mathbf{X} = \mathbf{x})}{\{\pi_1(\mathbf{Z}_0; \boldsymbol{\gamma}_0)f(l_1, l_2 \mid A^+ = 1, \mathbf{Z})\}^{A^+} \{1 - \pi_1(\mathbf{Z}_0; \boldsymbol{\gamma}_0)\}^{1-A^+}},$$

and $D(\mathbf{X}) = (1, A^+, A^+L_1, A^+L_2, A^+L_1L_2)'$.

Stabilized weights can be used to address extreme weights (Robins, 2000; Hernán and Robins, 2010), given by

$$w^s(\mathbf{X}, \mathbf{Z}) = \frac{\mathbb{I}(\mathbf{X} = \mathbf{x})\{\pi_2(\boldsymbol{\gamma}_0^\dagger)f(l_1, l_2 \mid A^+ = 1)\}^{A^+} \{1 - \pi_2(\boldsymbol{\gamma}_0^\dagger)\}^{1-A^+}}{\{\pi_1(\mathbf{Z}_0; \boldsymbol{\gamma}_0)f(l_1, l_2 \mid A^+ = 1, \mathbf{Z})\}^{A^+} \{1 - \pi_1(\mathbf{Z}_0; \boldsymbol{\gamma}_0)\}^{1-A^+}},$$

where $\pi_2(\boldsymbol{\gamma}_0^\dagger)$ and $f(l_1, l_2 \mid A^+ = 1)$ are unconditional quantities with respect to the conditional quantities $\pi_1(\mathbf{Z}_0; \boldsymbol{\gamma}_0)$ and $f(l_1, l_2 \mid A^+ = 1, \mathbf{Z})$ for weights stabilization. The inverse weighted estimating function for $\boldsymbol{\varphi}$ with stabilized weights is given by (2.24) with $w^s(\mathbf{X}, \mathbf{Z})$ in place of $w(\mathbf{X}, \mathbf{Z})$, that is

$$U_0(Y \mid \mathbf{X}; \boldsymbol{\varphi}, \boldsymbol{\Gamma}, \boldsymbol{\gamma}_0^\dagger, \boldsymbol{\gamma}_0^\dagger) = w^s(\mathbf{X}, \mathbf{Z})D(\mathbf{X}) \{Y - \mathbb{E}^R(Y \mid \mathbf{X}; \boldsymbol{\varphi})\}. \quad (2.25)$$

The estimate of $\boldsymbol{\varphi}$ is the solution from solving $U_0(Y \mid \mathbf{X}; \boldsymbol{\varphi}, \boldsymbol{\Gamma}, \boldsymbol{\gamma}_0) = 0$ or $U_0(Y \mid$

$\mathbf{X}; \boldsymbol{\varphi}, \boldsymbol{\Gamma}, \boldsymbol{\Gamma}^\dagger, \gamma_0, \gamma_0^\dagger) = 0$; See Appendix A.2 and A.3 for more details of parameter estimation through the second-order generalized estimating equations (GEE2) (Zhao and Prentice, 1990; Kalema et al, 2016) and derivation of the asymptotic covariance of the estimator.

2.4.3 Misspecification of Propensity Scores

Regression Adjustment and Inverse Density Weighting

If we treat the bivariate continuous or semi-continuous exposures as conditionally independent, the generalized propensity score can be adjusted similarly as in the regression model (2.13), that is,

$$\mathbb{E}(Y \mid \mathbf{X}, \tilde{\mathbf{S}}) = \mathbf{X}'\boldsymbol{\varphi} + \tilde{\mathbf{S}}'\boldsymbol{\psi}, \quad (2.26)$$

where $\mathbf{X} = (1, A_1, A_2, A_1 A_2)'$ contains continuous exposures, $\tilde{\mathbf{S}} = (\mathbb{E}(A_1 \mid \mathbf{Z}_1), \mathbb{E}(A_2 \mid \mathbf{Z}_2), \mathbb{E}(A_1 \mid \mathbf{Z}_1)\mathbb{E}(A_2 \mid \mathbf{Z}_2))'$. For semi-continuous exposures, we have $\mathbf{X} = (1, A^+, A^+ L_1, A^+ L_2, A^+ L_1 L_2)'$, and $\tilde{\mathbf{S}} = (\pi_1(\mathbf{Z}_0), \mathbb{E}(L_1 \mid \mathbf{Z}_0, \mathbf{Z}_1), \mathbb{E}(L_2 \mid \mathbf{Z}_0, \mathbf{Z}_2), \mathbb{E}(L_1 \mid \mathbf{Z}_0, \mathbf{Z}_1)\mathbb{E}(L_2 \mid \mathbf{Z}_0, \mathbf{Z}_2))'$ in (2.26).

When exposures are continuous, the inverse density of exposure weighted estimating equation used for exposure effects estimation is given by

$$U_{10}(\boldsymbol{\varphi}; \boldsymbol{\delta}) = \frac{\mathbb{I}(\mathbf{X} = \mathbf{x}) f(a_1) f(a_2)}{f(a_1 \mid \mathbf{Z}_1; \boldsymbol{\delta}) f(a_2 \mid \mathbf{Z}_2; \boldsymbol{\delta})} D(\mathbf{X}) \{Y - \mathbb{E}^R(Y \mid \mathbf{X}; \boldsymbol{\varphi})\} = 0, \quad (2.27)$$

and when exposures are semi-continuous, the inverse density of exposure weighted estimating equation is given by

$$w^s(\mathbf{X}, \mathbf{Z}) D(\mathbf{X}) \{Y - \mathbb{E}^R(Y \mid \mathbf{X}; \boldsymbol{\varphi})\} = 0, \quad (2.28)$$

where

$$w^s(\mathbf{X}, \mathbf{Z}) = \frac{\mathbb{I}(\mathbf{X} = \mathbf{x}) \{\pi_2(\gamma_0^\dagger) f(l_1 \mid A^+ = 1) f(l_2 \mid A^+ = 1)\}^{A^+} \{1 - \pi_2(\gamma_0^\dagger)\}^{1-A^+}}{\{\pi_1(\mathbf{Z}_0; \gamma_0) f(l_1 \mid A^+ = 1, \mathbf{Z}_1) f(l_2 \mid A^+ = 1, \mathbf{Z}_2)\}^{A^+} \{1 - \pi_1(\mathbf{Z}_0; \gamma_0)\}^{1-A^+}},$$

and $D(\mathbf{X}) = (1, A^+, A^+ L_1, A^+ L_2, A^+ L_1 L_2)'$.

When the dependence model assumes no role of confounders on the association between exposures, the regression model adjusted by the generalized propensity score is similar to

(2.15),

$$\mathbb{E}(Y \mid \mathbf{X}, \tilde{\mathbf{S}}) = \mathbf{X}'\boldsymbol{\varphi} + \tilde{\mathbf{S}}'\boldsymbol{\psi}, \quad (2.29)$$

where $\mathbf{X} = (1, A_1, A_2, A_1A_2)'$, $\tilde{\mathbf{S}} = (\mathbb{E}(A_1 \mid \mathbf{Z}_1), \mathbb{E}(A_2 \mid \mathbf{Z}_2), \mathbb{E}(A_1A_2 \mid \mathbf{Z}_1, \mathbf{Z}_2))'$. For semi-continuous exposures, we have $\mathbf{X} = (1, A^+, A^+L_1, A^+L_2, A^+L_1L_2)'$, $\tilde{\mathbf{S}} = (\pi_1(\mathbf{Z}_0), \mathbb{E}(L_1 \mid \mathbf{Z}_0, \mathbf{Z}_1), \mathbb{E}(L_2 \mid \mathbf{Z}_0, \mathbf{Z}_2), \mathbb{E}(L_1L_2 \mid \mathbf{Z}_0, \mathbf{Z}_1, \mathbf{Z}_2))'$.

The inverse density weighting approach using the misspecified joint density of exposures is to solve the following estimating equation for continuous exposures,

$$U_{10}(\boldsymbol{\varphi}; \boldsymbol{\delta}) = \frac{\mathbb{I}(\mathbf{X} = \mathbf{x}) f(a_1, a_2)}{f(a_1, a_2 \mid \mathbf{Z}_1, \mathbf{Z}_2; \boldsymbol{\delta})} D(\mathbf{X}) \{Y - \mathbb{E}^R(Y \mid \mathbf{X}; \boldsymbol{\varphi})\} = 0. \quad (2.30)$$

For the semi-continuous exposures, the inverse density of exposure weighted estimating equation is to solve (2.28) with

$$w^s(\mathbf{X}, \mathbf{Z}) = \frac{\mathbb{I}(\mathbf{X} = \mathbf{x}) \{\pi_2(\gamma_0^\dagger) f(l_1, l_2 \mid A^+ = 1)\}^{A^+} \{1 - \pi_2(\gamma_0^\dagger)\}^{1-A^+}}{\{\pi_1(\mathbf{Z}_0; \gamma_0) f(l_1, l_2 \mid A^+ = 1, \mathbf{Z}_1, \mathbf{Z}_2)\}^{A^+} \{1 - \pi_1(\mathbf{Z}_0; \gamma_0)\}^{1-A^+}}. \quad (2.31)$$

Empirical Bias and Efficiency

In this section, we report on simulation studies based on the random sample of size 500 with 1000 repetitions to estimate the exposure effect when exposures are continuous or semi-continuous. The estimated exposure effect through both regression adjustment and inverse density weighting approach would be assessed in terms of the relative bias, i.e., $(\hat{\boldsymbol{\varphi}} - \boldsymbol{\theta})/\boldsymbol{\theta}$.

When exposures in the true response model (2.1) are bivariate continuous, we let $A_k = \gamma_{k0} + \gamma_{k1}Z_k + R_k$, $k = 1, 2$, where

$$\mathbf{R} = (R_1, R_2)' \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right),$$

$\mathbf{Z} = (Z_1, Z_2, Z_3)' \sim \text{MVN}(\mathbf{0}, I_3)$, $I_3 = \text{diag}(1, 1, 1)$ is identity matrix with size 3, and $Z_k \perp R_k$. Additionally, we have $\log\{(1 + \rho)/(1 - \rho)\} = \gamma_{30} + \gamma_{31}Z_3$. We let $\gamma_{10} = 2$ and $\gamma_{11} = \log 2$ such that $\mathbb{E}(A_1) = 2$, let $\gamma_{20} = 0.5$ and $\gamma_{21} = \log 1.25$ such that $\mathbb{E}(A_2) = 0.5$. We let $\gamma_{31} \in [-5, 5]$ and solve for γ_{30} such that $\mathbb{E}(\rho) = 0.5$.

The relative biases of exposure effects estimated from (2.26), (2.27), (2.29) and (2.30)

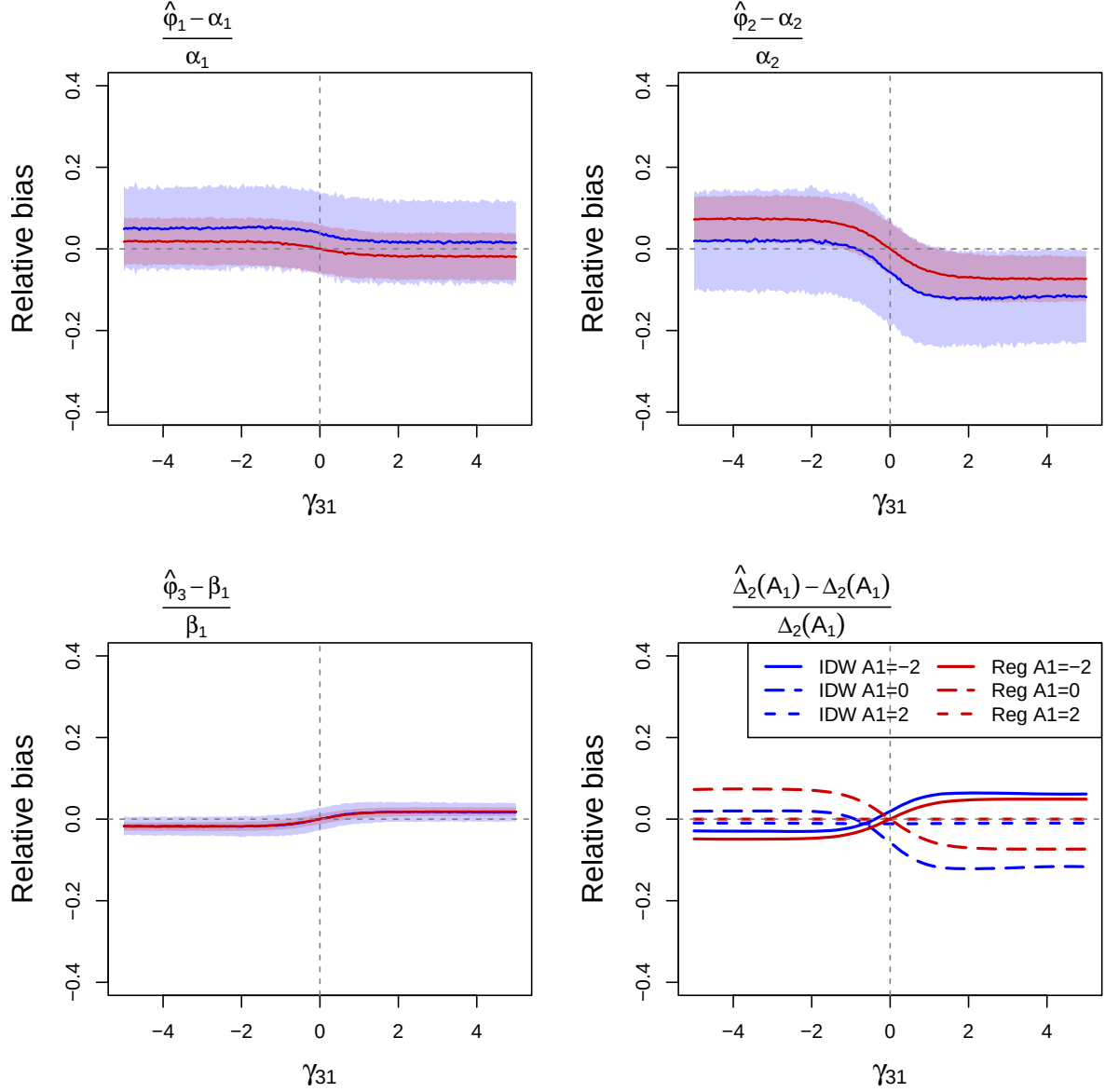


Figure 2.3: Empirical relative bias with standard errors (shaded bands) of estimates of exposure effects from inadequate propensity score modeling assuming conditional independence between two continuous exposures when the second-order confounder has the effect γ_{31} ; regression adjustment (Reg) (red) and inverse weighting (IPW) (blue) are considered; the marginal association between two continuous exposures is $\mathbb{E}(\rho) = 0.5$.

under two types of model misspecifications are shown in Figure 2.3 and Figure 2.4. The patterns of relative biases for continuous exposures are quite different from the case of binary exposures, but the main findings are similar in general. We again see the relative bias of the exposure effect increases as the covariate effect on the exposure dependence increases. Compared to the case of binary exposures, the estimated relative bias is relatively lower for continuous exposures. Additionally, the regression adjustment tends to produce a smaller bias than the inverse density weighting approach with higher efficiency.

For semi-continuous exposures, we let $\pi_1(\mathbf{Z}_0) = \text{expit}(\gamma_{00} + \gamma_{01}Z_1 + \gamma_{02}Z_2)$ where $\gamma_{00} = 1.484$, $\gamma_{01} = 1$, $\gamma_{02} = 2$ such that $\mathbb{E}\{\pi_1(\mathbf{Z}_0)\} = 0.7$ in the first part of the model. For the second part where given $A^+ = 1$, we suppose $\log\{A_k \mid A^+ = 1\} = \gamma_{k0} + \gamma_{k1}Z_k + R_k$, $k = 1, 2$, where

$$\mathbf{R} = (R_1, R_2)' \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right).$$

We let $\gamma_{10} = 2$ and $\gamma_{11} = \log 2$ such that $\mathbb{E}(\log A_1 \mid A^+ = 1) = 2$, let $\gamma_{20} = 0.5$ and $\gamma_{21} = \log 1.25$ such that $\mathbb{E}(\log A_2 \mid A^+ = 1) = 0.5$. Additionally, we have $\rho = \sigma_{12}/(\sigma_1\sigma_2)$, and $\log\{(1+\rho)/(1-\rho)\} = \gamma_{30} + \gamma_{31}Z_3$, where we let $\gamma_{31} \in [-5, 5]$ and solve for γ_{30} such that $\mathbb{E}(\rho) = 0.5$.

Figures A.1 and A.2 in Appendix A.4 show the relative biases of semi-continuous exposure effects estimated via (2.26), (2.28), (2.29) and (2.31) under two model misspecifications. The patterns of the relative bias as a function of the covariate effect γ_{31} are similar to those of Figures 2.3 and 2.4, while the case of semi-continuous exposures can lead to higher variability in estimates compared to binary and continuous exposures, especially under inverse density weighting and conditional independence. Again regression adjustment yields a smaller bias and greater precision compared to inverse density weighting in these simulation settings.

2.5 Two-Stage Causal Analysis for Bivariate Semi-Continuous Exposures

Here we propose a two-stage approach for causal analysis involving semi-continuous exposures. We again consider the special case that a single binary indicator A^+ indicates that an individual is exposed, and if $A^+ = 1$ then variables A_k record the different facets of exposure, $k = 1, \dots, p$. We let $A_k^c = A^+(A_k - \bar{A}_k)$ denote the centered value of A_k where $\bar{A}_k = \mathbb{E}(A_k \mid A^+ = 1)$, $k = 1, \dots, p$, and $\mathbf{A} = (A_1^c, \dots, A_p^c)'$ is the vector of exposures, $\mathbf{B} = (A_1^c A_2^c, \dots, A_{p-1}^c A_p^c)'$ is the vector of pairwise products of exposures,

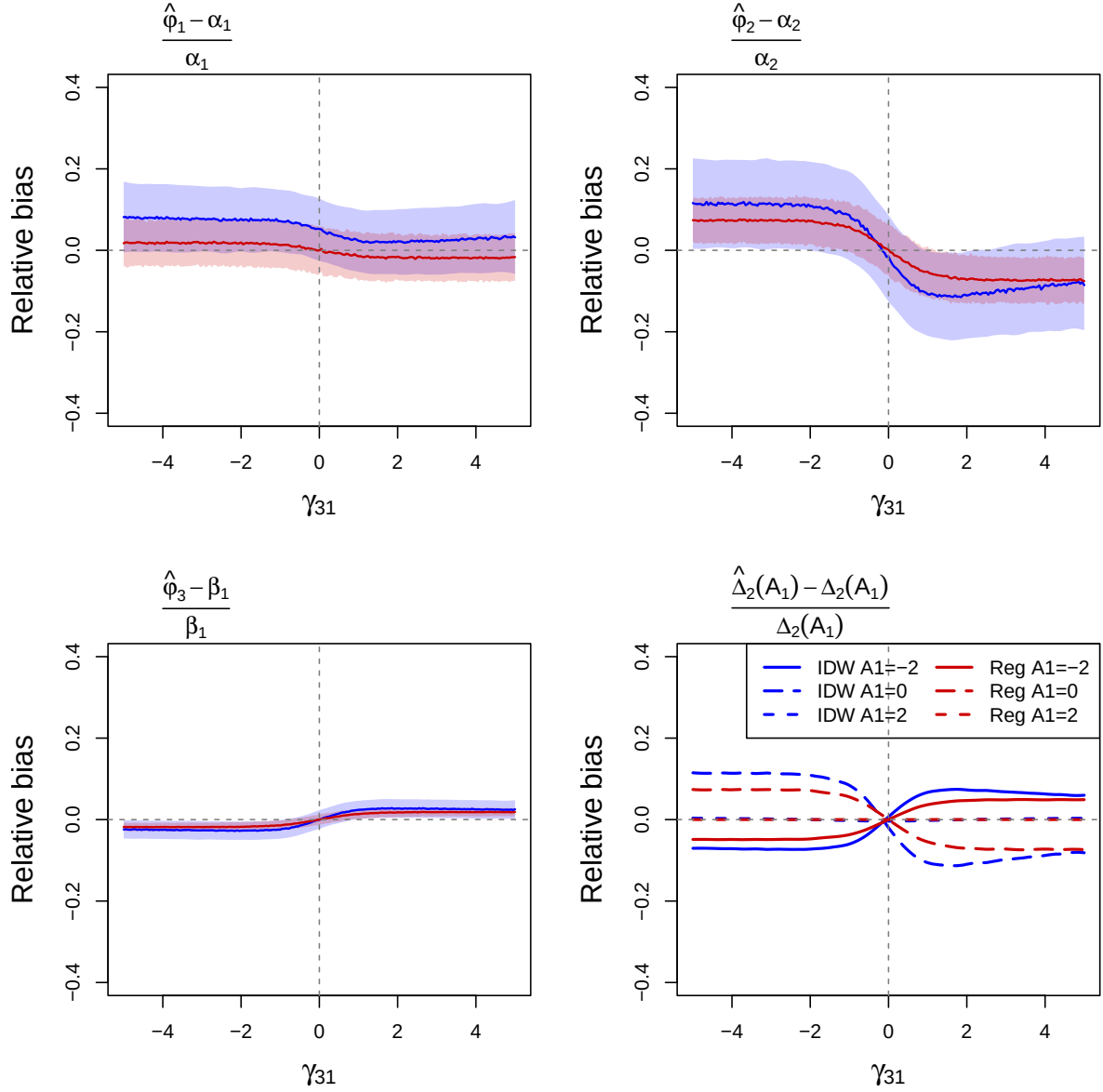


Figure 2.4: Empirical relative bias with standard errors (shaded bands) of estimates of exposure effects from inadequate propensity score modeling due to the omission of a second-order confounder with effect γ_{31} ; regression adjustment (Reg) (red) and inverse weighting (IPW) (blue) are considered; the marginal association between two continuous exposures is $\mathbb{E}(\rho) = 0.5$.

$\mathbf{X} = (A^+, \mathbf{A}', \mathbf{B}')$, and $\mathbf{Z}$ is an $r \times 1$ vector of confounders. The response model is

$$\mathbb{E}(Y \mid \mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}) = \beta_0 + \beta_1 A^+ + \boldsymbol{\beta}'_2 \mathbf{A} + \boldsymbol{\beta}'_3 \mathbf{B} + \boldsymbol{\beta}'_4 \mathbf{Z} . \quad (2.32)$$

2.5.1 Stage I: The Dose-Response Surface

In stage I we focus on the subgroup analysis modeling the dose-response surface among those with $A^+ = 1$ (the exposed). For a sample with $\{Y_i, \mathbf{X}_i, \mathbf{Z}_i, i = 1, \dots, n\}$, the estimating function is $U_1(\boldsymbol{\theta}) = \sum_{i=1}^n U_{i1}(\boldsymbol{\theta})$ for the covariate adjustment where

$$U_{i1}(\boldsymbol{\theta}) = A_i^+ D_1(\mathbf{A}_i, \mathbf{B}_i, \mathbf{Z}_i) \{Y_i - \eta_1(\mathbf{A}_i, \mathbf{B}_i, \mathbf{Z}_i; \boldsymbol{\theta})\} , \quad (2.33)$$

$D_1(\mathbf{A}_i, \mathbf{B}_i, \mathbf{Z}_i) = (1, \mathbf{A}'_i, \mathbf{B}'_i, \mathbf{Z}'_i)'$, and

$$\eta_1(\mathbf{A}_i, \mathbf{B}_i, \mathbf{Z}_i; \boldsymbol{\theta}) = \theta_1 + \boldsymbol{\theta}'_2 \mathbf{A}_i + \boldsymbol{\theta}'_3 \mathbf{B}_i + \boldsymbol{\theta}'_4 \mathbf{Z}_i .$$

Alternatively, if $\mathbf{S}_A = \mathbb{E}(\mathbf{A} \mid A^+ = 1, \mathbf{Z})$ and $\mathbf{S}_B = \mathbb{E}(\mathbf{B} \mid A^+ = 1, \mathbf{Z})$, we can set

$$U_{i2}(\boldsymbol{\theta}) = A_i^+ D_1(\mathbf{A}_i, \mathbf{B}_i, \mathbf{S}_i) \{Y_i - \eta_1(\mathbf{A}_i, \mathbf{B}_i, \mathbf{S}_i; \boldsymbol{\theta})\} , \quad (2.34)$$

where $D_1(\mathbf{A}_i, \mathbf{B}_i, \mathbf{S}_i) = (1, \mathbf{A}'_i, \mathbf{B}'_i, \mathbf{S}'_i)'$, and

$$\eta_1(\mathbf{A}_i, \mathbf{B}_i, \mathbf{S}_i; \boldsymbol{\theta}) = \theta_1 + \boldsymbol{\theta}'_2 \mathbf{A}_i + \boldsymbol{\theta}'_3 \mathbf{B}_i + \boldsymbol{\theta}'_4 \mathbf{S}_i .$$

We have $\hat{\boldsymbol{\theta}}$ from solving $\sum_{i=1}^n U_{i1}(\boldsymbol{\theta}) = 0$ or $\sum_{i=1}^n U_{i2}(\boldsymbol{\theta}) = 0$ which is the causal effect of the continuous part of exposure.

2.5.2 Stage II: The Effect of Exposure Status

The exposure effect of $\mathbf{A}$ and $\mathbf{B}$ is obtained from stage I and is used as an offset in stage II. Specifically we use $\tilde{\eta}_0(\mathbf{A}, \mathbf{B}; \boldsymbol{\theta}) = \hat{\boldsymbol{\theta}}'_2 \mathbf{A}_i + \hat{\boldsymbol{\theta}}'_3 \mathbf{B}_i$ as an offset taking the value zero for those with $A^+ = 0$. Stage II is directed at assessing the effect of the binary exposure A^+ ; many propensity score-based methods can be used to assess the effect of exposure at the reference level. We now discuss 4 possible approaches: simple regression adjustment, propensity score regression adjustment, inverse probability weighting (IPW), and augmented inverse probability weighting (AIPW).

A. SIMPLE REGRESSION ADJUSTMENT

For the simple regression adjustment we have the estimating function as

$$U_{i3}(\boldsymbol{\alpha}^\circ, \boldsymbol{\theta}) = D_2(A_i^+, \mathbf{Z}_i) \{Y_i - \eta_2(A_i^+, \mathbf{Z}_i; \boldsymbol{\alpha}^\circ, \boldsymbol{\theta})\} , \quad (2.35)$$

where $\boldsymbol{\alpha}^\circ = (\alpha_0, \alpha_1, \boldsymbol{\alpha}'_4)'$, $D_2(A_i^+, \mathbf{Z}_i) = (1, A_i^+, \mathbf{Z}'_i)'$, and

$$\eta_2(A_i^+, \mathbf{Z}_i; \boldsymbol{\alpha}^\circ, \boldsymbol{\theta}) = \alpha_0 + \alpha_1 A_i^+ + \boldsymbol{\alpha}'_4 \mathbf{Z}_i + \text{offset}(\tilde{\eta}_0^Z(\mathbf{A}, \mathbf{B}; \boldsymbol{\theta})) .$$

The superscript Z in $\tilde{\eta}_0^Z$ indicates that $\boldsymbol{\theta}$ is estimated from (2.33) by adjusting for $\mathbf{Z}$. Let $\Psi = (\boldsymbol{\alpha}^{\circ'}, \boldsymbol{\theta}')$, we solve the joint estimating equations

$$\sum_{i=1}^n U_i(\Psi) = \sum_{i=1}^n \begin{pmatrix} U_{i1}(\boldsymbol{\theta}) \\ U_{i3}(\boldsymbol{\alpha}^\circ, \boldsymbol{\theta}) \end{pmatrix} = \mathbf{0}$$

to get $\hat{\Psi}$.

B. PROPENSITY SCORE REGRESSION ADJUSTMENT

Let $S_0 = \mathbb{E}(A^+ | \mathbf{Z})$, replacing $\mathbf{Z}$ with S_0 in (2.35) we obtain the estimating functions

$$U_{i4}(\boldsymbol{\alpha}^\circ, \boldsymbol{\theta}) = D_2(A_i^+, S_{i0}) \{Y_i - \eta_2(A_i^+, S_{i0}; \boldsymbol{\alpha}^\circ, \boldsymbol{\theta})\} , \quad (2.36)$$

$$U_{i5}(\boldsymbol{\psi}) = D_3(\mathbf{Z}_i) \{A_i^+ - \mathbb{E}(A_i^+ | \mathbf{Z}_i; \boldsymbol{\psi})\} , \quad (2.37)$$

where $D_2(A_i^+, S_{i0}) = (1, A_i^+, S_{i0})'$, $D_3(\mathbf{Z}_i) = (1, \mathbf{Z}_i)'$, $\mathbb{E}(A_i^+ | \mathbf{Z}_i; \boldsymbol{\psi}) = \text{expit}(\boldsymbol{\psi}' \mathbf{Z}_i)$, and

$$\eta_2(A_i^+, S_{i0}; \boldsymbol{\alpha}^\circ, \boldsymbol{\theta}) = \alpha_0 + \alpha_1 A_i^+ + \alpha_4 S_{i0} + \text{offset}(\tilde{\eta}_0^S(\mathbf{A}, \mathbf{B}; \boldsymbol{\theta})) .$$

The superscript S in $\tilde{\eta}_0^S$ indicates that $\boldsymbol{\theta}$ is estimated from (2.34) by adjusting with the generalized propensity score. Let $\Psi = (\boldsymbol{\alpha}^{\circ'}, \boldsymbol{\theta}', \boldsymbol{\psi}')$, we can consider solving the joint estimating equations to obtain $\hat{\Psi}$,

$$\sum_{i=1}^n U_i(\Psi) = \sum_{i=1}^n \begin{pmatrix} U_{i2}(\boldsymbol{\theta}) \\ U_{i4}(\boldsymbol{\alpha}^\circ, \boldsymbol{\theta}) \\ U_{i5}(\boldsymbol{\psi}) \end{pmatrix} = \mathbf{0} .$$

C. INVERSE PROBABILITY WEIGHTING (IPW)

We first consider a marginal structural model with the offset,

$$\eta_3(A^+; \phi, \theta) = \mathbb{E}(Y \mid A^+; \phi, \theta) = \phi_0 + \phi_1 A^+ + \text{offset}(\tilde{\eta}_0^S(\mathbf{A}, \mathbf{B}; \theta)) .$$

Then we define the estimating function

$$U_{i6}(\phi, \theta, \psi) = \frac{I(A_i^+ = a_i^+)}{P(A_i^+ = a_i^+ \mid \mathbf{Z}; \psi)} D_4(A_i^+) \{Y_i - \eta_3(A_i^+; \phi, \theta)\} ,$$

where $D_4(A_i^+) = (1, A_i^+)'$. Let $\Psi = (\phi', \theta', \psi')'$, so we can consider solving the joint estimating equations,

$$\sum_{i=1}^n U_i(\Psi) = \sum_{i=1}^n \begin{pmatrix} U_{i2}(\theta) \\ U_{i5}(\psi) \\ U_{i6}(\phi, \theta, \psi) \end{pmatrix} = \mathbf{0} .$$

D. AUGMENTED INVERSE PROBABILITY WEIGHTING (AIPW)

The AIPW approach combines an outcome regression (OR) model and a propensity score (PS) model. The AIPW estimators are shown to be doubly robust, with estimators consistent if either the outcome regression model or the propensity score model is correctly specified ([Scharfstein et al, 1999](#)). To obtain this estimator, we begin by specifying the estimating functions for the outcome regression model through the G-computation approach ([Snowden et al, 2011](#); [Williamson et al, 2012](#)).

We consider

$$\begin{aligned} U_{i7}(\theta, \lambda_1) &= A_i^+ D_3(\mathbf{Z}_i) \{Y_i - \lambda_1' \mathbf{Z} - \text{offset}(\tilde{\eta}_0^S(\mathbf{A}, \mathbf{B}; \theta))\} , \\ U_{i8}(\theta, \lambda_2) &= (1 - A_i^+) D_3(\mathbf{Z}_i) \{Y_i - \lambda_2' \mathbf{Z} - \text{offset}(\tilde{\eta}_0^S(\mathbf{A}, \mathbf{B}; \theta))\} . \end{aligned}$$

Then the AIPW estimator $\hat{\mu}_1$ and $\hat{\mu}_0$ can be obtained by solving the estimating equation $\sum_{i=1}^n U_{i9}(\theta, \psi, \mu_1) = 0$ and $\sum_{i=1}^n U_{i10}(\theta, \psi, \mu_0) = 0$, where

$$U_{i9}(\theta, \psi, \mu_1) = \frac{Y_i A_i^+ - \{A_i^+ - P(A_i^+ = 1 \mid \mathbf{Z}; \psi)\} \{\lambda_1' \mathbf{Z} + \text{offset}(\tilde{\eta}_0^S(\mathbf{A}, \mathbf{B}; \theta))\}}{P(A_i^+ = 1 \mid \mathbf{Z}; \psi)} - \mu_1 ,$$

and

$$U_{i10}(\boldsymbol{\theta}, \boldsymbol{\psi}, \mu_0) = \frac{Y_i(1 - A_i^+) - \{P(A_i^+ = 1 \mid \mathbf{Z}; \boldsymbol{\psi}) - A_i^+\} \{\boldsymbol{\lambda}'_2 \mathbf{Z} + \text{offset}(\tilde{\eta}_0^S(\mathbf{A}, \mathbf{B}; \boldsymbol{\theta}))\}}{1 - P(A_i^+ = 1 \mid \mathbf{Z}; \boldsymbol{\psi})} - \mu_0 .$$

Let $\Psi = (\boldsymbol{\theta}', \boldsymbol{\psi}', \boldsymbol{\lambda}', \boldsymbol{\mu}')'$, we can solve the joint estimating equations $\sum_{i=1}^n U_i(\Psi) = \mathbf{0}$, where

$$U_i(\Psi) = (U'_{i2}(\boldsymbol{\theta}), U'_{i5}(\boldsymbol{\psi}), U'_{i7}(\boldsymbol{\theta}, \boldsymbol{\lambda}_1), U'_{i8}(\boldsymbol{\theta}, \boldsymbol{\lambda}_2), U'_{i9}(\boldsymbol{\theta}, \boldsymbol{\psi}, \mu_1), U'_{i10}(\boldsymbol{\theta}, \boldsymbol{\psi}, \mu_2))' .$$

From standard estimating function theory, we have

$$\text{asvar}(\sqrt{n}(\hat{\Psi} - \Psi)) = \mathcal{A}^{-1}(\Psi) \mathcal{B}(\Psi) \{\mathcal{A}^{-1}(\Psi)\}' ,$$

where $\mathcal{A}(\Psi) = E\{-\partial U_i(\Psi)/\partial \Psi'\}$ and $\mathcal{B}(\Psi) = E\{U_i(\Psi)U'_i(\Psi)\}$, and the expectation is taken with respect to the true distribution. In practice, we use sample averages to estimate $\mathcal{A}(\Psi)$ and $\mathcal{B}(\Psi)$, and evaluate the resulting expressions at the estimates to obtain $\mathcal{A}^{-1}(\hat{\Psi})\mathcal{B}(\hat{\Psi})\{\mathcal{A}^{-1}(\hat{\Psi})\}'$, where

$$\mathcal{A}(\hat{\Psi}) = -\frac{1}{n} \sum_{i=1}^n \partial U_i(\Psi)/\partial \Psi' \Big|_{\Psi=\hat{\Psi}} ,$$

and

$$\mathcal{B}(\hat{\Psi}) = \frac{1}{n} \sum_{i=1}^n U_i(\Psi)U'_i(\Psi) \Big|_{\Psi=\hat{\Psi}} .$$

2.5.3 Simulation Studies on Two-Stage Approach

Here we evaluate the performance of the two-stage approach for analysis of a bivariate semi-continuous exposure under the data-generating process of Section 2.4.3. We consider the stage II estimator based on various combinations of methods in stages I and II. For stage I, confounder regression (2.33) and propensity score adjustment (2.34) are used. In stage II we employ methods A) to D) of Section 2.5.2. We see that when all models are correctly specified, the empirical bias is negligible across all methods as shown in Table 2.1; use of IPW in stage II yields the estimates with the highest standard error, while the AIPW estimator improves the efficiency as expected.

Table 2.2 displays the relative bias and empirical standard error of the stage II estimators when models are misspecified in stage II. Here the confounders are transformed

Table 2.1: Simulation results with sample size $n = 500$, repetitions $n_{sim} = 1000$: relative bias and standard error of stage II estimator of the binary exposure effect when models are correctly specified.

Stage I	Stage II	Bias (%)	Empirical SE
Regression with Z	Regression with Z	-0.08	0.14
Regression with GPS	Regression with Z	-0.12	0.15
	Regression with PS	0.13	0.19
	IPW	0.16	0.23
	AIPW	-0.11	0.17

as $V_1 = \exp(Z_1/2) + Z_2$, $V_2 = Z_2/(1 + \exp(Z_1)) + 10$, $V_3 = (Z_1 Z_2/10 + 0.6)^2$ and $\mathbf{V} = (V_1, V_2, V_3)'$ is used instead of $\mathbf{Z} = (Z_1, Z_2, Z_3)'$ in stage II for model misspecification. If either the outcome regression (OR) model or the propensity score (PS) model is misspecified in Stage II, biased estimates are obtained in methods reliant on the respective model. The AIPW estimator exhibits the double robustness property. When both models are misspecified in stage II, the AIPW approach yields an estimator with a large finite sample bias and high standard error.

Table 2.2: Simulation results with sample size $n = 500$, repetitions $n_{sim} = 1000$: relative bias and standard error of stage II estimator of the binary exposure effect with model misspecification in stage II.

Model misspecification	Stage I	Stage II	Bias (%)	Empirical SE
PS model	Regression with GPS	Regression with PS	0.91	0.18
		IPW	-5.06	0.35
		AIPW	0.07	0.31
OR model	Regression with Z	Regression with Z	4.48	0.18
	Regression with GPS	Regression with Z	4.45	0.18
		AIPW	-0.12	0.22
PS+OR model	Regression with GPS	AIPW	5.95	0.43

2.6 Prenatal Alcohol Exposure and Cognition

Here we consider a large prospective cohort study comprised of 480 inner-city pregnant African-American mothers recruited between September 1986 and April 1989 (Jacobson et al, 2004). Quantitative measures of alcohol consumption during pregnancy were obtained at each maternal prenatal visit using a timeline follow-back interview (Jacobson et al, 2002) and converted to three common metrics: the number of standard drinks per drinking occasion, frequency of drinking (days/month), and average oz absolute alcohol (AA/day). Additional variables collected included socioeconomic status (SES), maternal age, maternal years of education, marital status, and family income; maternal smoking (cigarettes/day) and marijuana, cocaine, and other illicit drug use (days/month), maternal history of alcohol abuse or dependence; and child’s sex, race, and whether s/he was raised by his/her biological mother and/or was in foster care; see Table A.1 in Appendix A.4. The mothers and their children were followed for 19-21 years postpartum with a retention rate of 79.0% ($N = 377$). A variety of cognitive tests were administered to each child to assess various facets of cognitive function. The results of the various cognitive tests were synthesized by fitting a second-order confirmatory factor analysis giving a composite measure of cognitive function (Jacobson et al, 2024) which we adopted as the response, Y , in the analyses that follow.

Among the 377 mothers in this analysis, 16.2% reported no alcohol consumption during pregnancy resulting in the alcohol exposure variables being semi-continuous. For mothers who drank alcohol during pregnancy, two main exposures (oz prenatal absolute alcohol exposure per drinking occasion, and the proportion of drinking days during pregnancy) were log-transformed and centered. Here we evaluate the causal effects in four ways as discussed in the previous sections: subgroup analysis involving drinkers, analyses based on the bivariate semi-continuous exposure using conventional regression adjustment and the two-part model, and full sample analysis with the two-stage approach. We aim to evaluate the causal effect of alcohol drinking, two main exposures, and their interaction on child cognitive function while controlling for selected measurements listed in Table 2.3.

2.6.1 Subgroup Analysis involving Drinkers

We first restrict attention to drinkers and focus on the causal effects of alcohol exposure on children’s cognition. Let A_1 represent the number of standard drinks per drinking occasion, and A_2 represent the proportion of days in each month of drinking (days/month); we refer to these as the intensity and frequency of drinking respectively. If $A^+ = \mathbb{I}(A_{i1}A_{i2} > 0)$

Table 2.3: Estimates of the effects for each confounder.

	Estimate (SE)			
	logit(π_{1i})	$\mathbb{E}(L_{1i} \mid A_i^+ = 1)$	$\mathbb{E}(L_{2i} \mid A_i^+ = 1)$	$\log \frac{(1+\rho_i)}{(1-\rho_i)}$
Intercept	-0.625 (1.794)	-1.499 (0.613)	-0.648 (0.636)	-0.184 (8.978)
Smoking during pregnancy (cigarettes/day)	0.090 (0.025)	0.007 (0.005)	0.009 (0.005)	0.018 (0.108)
Socioeconomic status at birth/infancy	0.012 (0.020)	-0.016 (0.007)	-0.001 (0.007)	0.040 (0.101)
Child's gender (female)	-0.574 (0.310)	-0.042 (0.102)	-0.075 (0.105)	0.297 (2.169)
Mother's age at delivery	0.048 (0.035)	0.020 (0.011)	0.036 (0.011)	-0.037 (0.315)
Prenatal Beck Depression Inventory	0.024 (0.022)	0.011 (0.007)	0.012 (0.007)	0.004 (0.168)
HOME score at infancy	0.023 (0.033)	0.003 (0.011)	-0.013 (0.012)	-0.023 (0.241)
Biological mother's verbal IQ (PPVT score)	0.000 (0.013)	-0.003 (0.004)	0.000 (0.004)	0.020 (0.086)
Biological mother's marital status (married)	-0.121 (0.498)	-0.001 (0.179)	0.098 (0.186)	-0.350 (3.102)
Prenatal cocaine exposure (days/month)	-0.140 (0.061)	-0.026 (0.019)	-0.031 (0.020)	0.828 (3.025)
Prenatal marijuana exposure (days/month)	0.226 (0.137)	-0.016 (0.019)	0.021 (0.020)	-0.002 (0.198)
Prenatal opiate exposure (days/month)	0.122 (0.234)	0.071 (0.044)	0.163 (0.046)	-0.010 (0.785)
Number of prenatal visits	-0.018 (0.059)	-0.002 (0.020)	-0.111 (0.021)	-0.076 (0.455)
Biological mother's education (year)	-0.040 (0.116)	0.092 (0.038)	0.012 (0.040)	-0.011 (0.575)
Parity	-0.474 (0.213)	0.095 (0.059)	-0.082 (0.062)	-0.134 (1.357)
Gravidity	0.288 (0.144)	-0.019 (0.036)	0.040 (0.037)	0.064 (0.660)
Gestational age at screening	-0.011 (0.025)	0.007 (0.008)	0.014 (0.008)	0.001 (0.153)

indicates any consumption of alcohol, to conduct the analyses among drinkers we consider the regression model

$$\mathbb{E}(Y_i \mid A_i^+ = 1, \mathbf{L}_i, \mathbf{Z}_i) = \mu + \alpha_1 L_{i1} + \alpha_2 L_{i2} + \alpha_3 L_{i1} L_{i2} + \mathbf{Z}_i' \boldsymbol{\zeta}, \quad (2.38)$$

where $L_{ik} = A_i^+ \{\log A_{ik} - \mathbb{E}\{\log A_{ik} \mid A_i^+ = 1\}\}$. When using propensity score regression adjustment we consider

$$\mathbb{E}(Y_i \mid A_i^+ = 1, \mathbf{L}_i, \mathbf{S}_i) = \mu + \alpha_1 L_{i1} + \alpha_2 L_{i2} + \alpha_3 L_{i1} L_{i2} + \mathbf{S}_i' \boldsymbol{\zeta}, \quad (2.39)$$

where $\mathbf{S}_i = (S_{i1}, S_{i2}, S_{i3})'$, with $S_{ik} = \mathbb{E}(L_{ik} \mid A_i^+ = 1, \mathbf{Z}_i)$, $k = 1, 2$, $S_{i3} = \mathbb{E}(L_{i1} L_{i2} \mid A_i^+ = 1, \mathbf{Z}_i)$. The estimated effects of each confounder are given in columns 3-5 of Table 2.3 for each of the respective models.

2.6.2 Analyses based on Bivariate Semi-Continuous Exposure Model

The conventional regression model adjusted by all the confounders is fitted as

$$\mathbb{E}(Y_i | A_i^+, \mathbf{L}_i, \mathbf{Z}_i) = \nu + \lambda_0 A_i^+ + \lambda_1 L_{i1} + \lambda_2 L_{i2} + \lambda_3 L_{i1} L_{i2} + \mathbf{Z}_i' \boldsymbol{\eta},$$

with $\mathbf{Z}_i$ the set of confounders. Using a two-part model for the generalized propensity score as in [Akkaya Hocagil et al \(2021\)](#), the regression model adjusted by generalized propensity score is

$$\mathbb{E}(Y_i | A_i^+, \mathbf{L}_i, \mathbf{S}_i) = \nu + \lambda_0 A_i^+ + \lambda_1 L_{i1} + \lambda_2 L_{i2} + \lambda_3 L_{i1} L_{i2} + \mathbf{S}_i' \boldsymbol{\eta},$$

where as in [\(2.23\)](#), $\mathbf{S}_i = (\pi_{i1}(\mathbf{Z}_i), \mathbb{E}(L_{i1} | \mathbf{Z}_i), \mathbb{E}(L_{i2} | \mathbf{Z}_i), \mathbb{E}(L_{i1} L_{i2} | \mathbf{Z}_i))'$ with

$$\begin{aligned} \pi_{i1}(\mathbf{Z}_i) &= \text{expit}(\mathbf{Z}_i' \boldsymbol{\gamma}_0), \\ \mathbb{E}(L_{ik} | \mathbf{Z}_i) &= \pi_{i1}(\mathbf{Z}_i) \mathbb{E}(L_{ik} | A_i^+ = 1, \mathbf{Z}_i), k = 1, 2, \\ \mathbb{E}(L_{i1} L_{i2} | \mathbf{Z}_i) &= \pi_{i1}(\mathbf{Z}_i) \mathbb{E}(L_{i1} L_{i2} | A_i^+ = 1, \mathbf{Z}_i). \end{aligned}$$

The results are shown in [Table 2.3](#).

The inverse density weighting approach to estimating causal effects involves solving [\(2.25\)](#) with stabilized weights. We assume the exposures follow a bivariate normal distribution when we estimate the conditional joint density of the two exposures.

2.6.3 Full Sample Analysis with Two-Stage Approach

We also address the effects of confounding variables through the two-stage approach as discussed in [Section 2.5](#). We first define the offset from subgroup analysis on drinkers using estimates obtained from [\(2.38\)](#) or [\(2.39\)](#) which is the stage I. Let $\tilde{\eta}_i = \hat{\alpha}_1 L_{i1} + \hat{\alpha}_2 L_{i2} + \hat{\alpha}_3 L_{i1} L_{i2}$ be the offset in stage II and use regression adjustment with covariates, regression adjustment with propensity score, IPW, and AIPW to differentiate the actual exposure effect as a drinker.

2.6.4 Results

We assess the covariate balance using Pearson correlation ([Austin, 2019](#); [Williams and Crespi, 2020](#)); see [Table A.2](#) in [Appendix A.4](#). The Pearson correlations were computed between each exposure and each covariate in [Table 2.3](#) before generalized propensity score

adjustment or weighting adjustment. To evaluate the performance of each approach, we analyzed the covariate balance after conditioning on the generalized propensity score or applying inverse probability/density weighting. All the methods described above substantially reduce the maximum absolute correlation and average absolute correlation compared to the conventional regression approach shown in the third row of Table A.2, indicating improvement of covariate balance.

In the full sample analysis with the two-part model, we found that the regression approach with generalized propensity score adjustment reduced covariate imbalance more effectively than the inverse density weighting approach. In the full sample analysis with the two-stage approach, the propensity score-based methods in stage II (e.g., IPW) achieve better covariate balance than the regression approach. Overall, our results demonstrate the effectiveness of each analysis with different methods in improving covariate balance, with the optimal approach depending on the specific analysis context.

Table 2.4: Estimated effects of being a drinker versus non-drinker, the logarithm of prenatal absolute alcohol exposure per occasion, the logarithm of proportion of drinking days across pregnancy, and their interaction.

Analysis	Method	Average causal effect (SE)			
		Drinker indicator	Alcohol per occasion	Proportion drinking days	Interaction
Subgroup analysis	Covariate adjustment	-	-1.80 (1.07)	-0.50 (0.94)	-1.39 (0.85)
	GPS adjustment	-	-1.80 (1.09)	-0.50 (0.96)	-1.39 (0.86)
Full sample analysis with conventional regression	Covariate adjustment	-0.41 (2.08)	-2.09 (1.05)	-0.88 (0.91)	-1.62 (0.83)
Full sample analysis with two-part model	GPS adjustment	-1.26 (2.22)	-1.84 (1.11)	-0.93 (0.95)	-1.39 (0.90)
	IDW	-1.43 (2.10)	-2.22 (1.18)	-0.14 (0.96)	-0.44 (0.84)
Full sample analysis with two-stage approach [†]	Reg Z + Reg Z	-0.35 (1.48)	-	-	-
	Reg GPS + Reg Z	-0.35 (1.96)	-	-	-
	Reg GPS + Reg PS	-2.90 (2.29)	-	-	-
	Reg GPS + IPW	-3.81 (2.35)	-	-	-
	Reg GPS + AIPW	-4.34 (1.98)	-	-	-

[†] Reg Z = Regression with covariate adjustment; Reg GPS = Regression with generalized propensity score adjustment; Reg PS = Regression with propensity score adjustment.

Table 2.4 presents the causal effects estimated from both subgroup analysis and full sample analysis with the various methods. All estimated effects are negative, including the effects of drinking at the reference level versus not drinking, prenatal absolute alcohol exposure per occasion, the proportion of drinking days, and their interaction, confirming that alcohol consumption by expectant mothers has a negative effect on children’s cognitive development. For subgroup analysis among drinkers, the results from the regression with covariate adjustment and the regression model adjusted by generalized propensity scores are quite similar, indicating the cognitive function score of the child decreases as the mother increases the amount of alcohol consumed at each drinking occasion, or increases her drinking frequency. The interaction term is negative so the effect of increasing the intensity of drinking on each occasion is greater when the frequency of alcohol use during pregnancy is greater, and vice versa. For full sample analysis, the conventional regression adjustment and the regression model adjusted by the two-part generalized propensity scores also generate similar estimates except for the drinker indicator. The inverse density weighting approach results in a larger effect size of alcohol exposure per occasion and smaller effect sizes of drinking frequency and their interaction compared with estimates obtained from the regression methods. In full analysis with the two-stage approach, the effect of drinking obtained from the propensity score based methods in stage II, that is regression with a propensity score adjustment, IPW, and AIPW method, is larger than the estimate obtained from regression approaches. The IPW approach in stage II has the largest standard error among all the methods, whereas the AIPW approach generates a similar, but more precise estimate of the causal effect.

We built a dose-response surface on the exposure estimates obtained from stage I using a regression model with generalized propensity score adjustment in the two-stage approach. The resulting surface is shown in the left panel of Figure 2.5. We define the reference level of a child’s cognition when the mother has an average level of intensity and density of alcohol use, that is having 1.81 oz absolute alcohol per occasion and spending approximately 4 days per month (0.13 months) consuming alcohol. The x -axis and y -axis show the increment levels of drinking frequency and intensity respectively. This graphical depiction shows the expected reduction in child’s cognitive function resulting from increased drinking frequency and intensity. The right panel of Figure 2.5 displays a contour plot that shows the change in cognition associated with different combinations of changes in intensity and frequency of alcohol intake. For instance, spending approximately 9 more days per month (0.3 months) drinking based on the reference level while maintaining the same drinking intensity, or spending 3 more days per month while also increasing the dose of drinks by 2 oz on each drinking occasion based on the reference level would both result in a 4-unit decrease in the child’s cognitive function score.

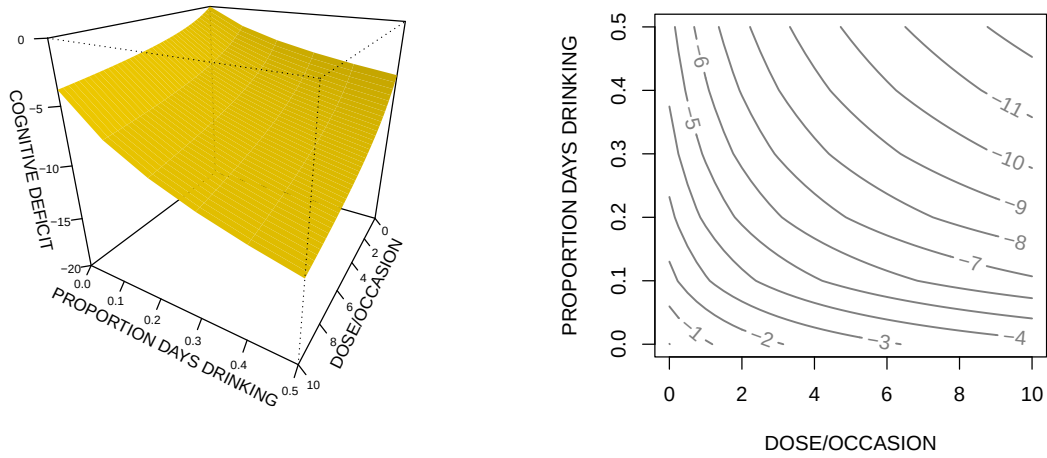


Figure 2.5: Dose-response surface and contour plot generated from stage I in two-stage analysis: the effect of changes in drinking frequency (proportion days drinking) and intensity (dose/occasion) on child cognitive deficit; reference levels: 1.81 oz AA/occasion, 0.13 of the pregnancy drinking.

2.7 Remarks

In this chapter we describe methods for assessing the effect of two or more exposure variables on a continuous response. Confounding is addressed through the use of a generalized propensity score. In this setting the generalized propensity score is based on the joint density of exposures and in particular is the vector of linear functions of confounders necessary to construct this joint density. For inverse density weighting we stabilize the weights via the marginal joint density of the exposures. A key finding is that when confounders affect the joint distribution of the exposure variables they must be accounted for in the propensity score modeling to ensure consistent estimation and valid inference regarding the interactions between the exposures. Regression adjustment appears to be less sensitive to misspecification of the dependence structure and omission of the second-order effects of confounders, and moreover leads to more precise estimation of causal effects. We did not consider matching or stratification on the generalized propensity score but these represent alternative approaches for dealing with confounders (Wu et al, 2022).

Note that much of the literature on causal analysis includes the conceptualization of potential outcomes (Splawa-Neyman et al, 1990; Rubin, 1974) for each unit, here denoted by $\{Y(\mathbf{a}) \text{ for } \mathbf{a} \in \mathcal{A}\}$ where $\mathcal{A}$ may be a multidimensional space for possible exposures. We do not make use of potential outcomes in this chapter, but our methods could be reformulated and formalized within that framework.

Chapter 3

The Design of Large-Scale Plasma Donation Trials

3.1 Introduction

The need for blood plasma is expected to increase in coming years ([Sanchez et al, 2001](#)) with machine learning algorithms ([Xiang et al, 2021](#); [Li et al, 2023](#)) now being used for demand forecasting. Acute periods of need can increase the intensity of recruitment efforts, as was seen in the early period of the COVID-19 pandemic when convalescent plasma was needed for clinical trials evaluating its therapeutic effect at a time when treatment options were limited ([Bégin et al, 2021](#)). Forecasting need is only helpful if mitigation strategies are available to avoid crises from shortfalls. Campaigns are effective ([Berger, 2013](#)) at increasing the pool of blood donors and can help encourage established donors to increase their donation frequency in compliance with best practice guidelines. More generally, efforts to increase the efficiency ([Collins et al, 2015](#)) of the blood supply system can ensure products are used in an effective manner and reduce waste ([Rahdar et al, 2018](#)). Shortfalls in plasma supply can also be addressed by drawing more plasma on donation occasions, and new algorithms have been developed to express the relation between the volume of plasma withdrawn from donors and risk of hypotensive adverse events. Such algorithms can inform the development of nomograms designed to individualize the target volume of plasma based on sex, body size, donation history, and other factors ([Zhao et al, 2023](#)).

It is important to establish the safety of any proposed nomograms. [Hartmann et al \(2021\)](#) conducted a non-inferiority trial of such a new algorithm with the goal of showing that the personalized volume collected in donors did not lead to an unacceptably higher

rate of significant hypotensive (e.g. vasovagal or hypovolemia) events. The results showed that the new nomogram led to an average of an 8% greater volume of plasma per donation while not affecting the rate of significant hypotensive adverse events. The design of this study was based on an extensive Monte Carlo simulation study using historical data on donation frequency and rates of significant hypertensive adverse events.

We consider an innovative and rigorous design of such a trial aiming to demonstrate the non-inferiority of a new algorithm for determining the plasma volume to withdrawn with respect to hypotensive adverse events. For some higher risk donors a lower target may be set but for others a higher volume may be set with the potential that across the whole donor population more plasma can be extracted. We suppose individuals are randomized to donate plasma according to the standard algorithm or a new algorithm. As in [Hartmann et al \(2021\)](#) we define the primary response as an indicator that a serious hypotensive adverse event occurring following a donation. Trial participants will donate plasma several times during the study period, and the number of donations will vary appreciably across individuals due to inter-individual variation in the propensity to donate. Moreover, because of inter-individual variation in risk, the responses to successive donations from a given donor cannot be treated as independent. The resulting data is therefore serially clustered (within donor) longitudinal binary data with a cluster-level covariate indicating the arm to which the donor is randomized. Since the event rate is very low in this setting a Poisson approximation to the Bernoulli distribution can be adopted and we formulate a regression model based on the log link and identity variance function. A robust variance is used both to recognize that the variance function is misspecified and to account for the within-individual dependence in the response to successive donations. The robust variance formula requires calculation under model assumptions at the design stage, and we use the corresponding asymptotic variance to plan the the duration of the accrual and treatment periods.

The remainder of this Chapter is organized as follows. In [Section 3.2](#) we define notation, specify the model, give the generalized estimating functions used for estimation, and present the large sample theory to guide the trial design. In [Section 3.3](#) we focus on a particular setting in detail involving a simple marginal model for treatment effect. The precise form of the large sample distribution is given explicitly for this simplified model, which makes it clear which terms are needed to facilitate planning of a two-arm clinical trial. One element of this large sample distribution involves the distribution of the number of donations during the follow-up period and the other is the joint distribution of the responses over successive donations. We derive the distribution of the number of donations in the follow-up phase of the study conditioning on the number of donations in the recruitment phase being zero-truncated. The joint distribution of responses is formulated

in terms of a multivariate binary response model and we use a log link for the marginal model expressing the relation between the risk of an event as a function of the arm of the trial. The serial dependence in the responses is assumed to be exchangeable in the data generation process but a working independence assumption is used in the generalized estimating functions. In Section 3.4 we consider study design for both superiority and non-inferiority trials and show empirically that the methods of analysis maintain control over the type I error and the sample size formula ensures the desired statistical power is achieved. With a fixed study period (over which recruited donors provide donations), the size of the study donor population determines how long the recruitment period is expected to be to ensure adequate power. We show that the design maintains good power and control of the type I error rate whether the number of donors is treated as random or whether the trial specifies the expected number of recruited individuals as the required sample size (fixing the sample size and letting the duration of the accrual period be random). In Section 3.4.3 we illustrate an example of a plasma donation study design with different donor population sizes and investigate the impact of misspecification of the covariate effect and the correlation of responses within individuals on study power. Remarks are provided in Section 3.5. Group sequential analysis with the interim analyses planned over the course of the trials is discussed in Section 3.6 to 3.12.

3.2 Statistical Formulation and Methods

3.2.1 Notation and Generalized Estimating Equations

Consider a sample of n independent individuals recruited to a two-armed donation trial and let X_i indicate that individual i was assigned to experimental arm with $X_i = 0$ otherwise, $i = 1, \dots, n$. Suppose donor i makes K_i donations over the course of follow-up and let Y_{ik} indicate the occurrence of an adverse event upon the k th donation (i.e. $Y_{ik} = 1$ if an adverse event occurred and $Y_{ik} = 0$ otherwise), $k = 1, \dots, K_i$, $i = 1, \dots, n$. The vector of indicators $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iK_i})'$ denotes the full set of K_i measurements for individual i , $i = 1, \dots, n$. If $\mathbf{V}_{ik}$ is a vector of auxiliary covariates we let $\mathbf{Z}_{ik} = (1, X_i, \mathbf{V}_{ik}')'$ and define $\mu(\mathbf{Z}_{ik}; \boldsymbol{\beta}) = \mathbb{E}(Y_{ik} \mid \mathbf{Z}_{ik}; \boldsymbol{\beta}) = P(Y_{ik} = 1 \mid \mathbf{Z}_{ik}; \boldsymbol{\beta})$ as the conditional expectation of Y_{ik} given covariates where $\boldsymbol{\beta}$ indicates the effects of covariates with p dimensions.

For regression models with binary responses we typically use either the logit (i.e. $\log(\mu/(1 - \mu))$), probit (i.e. $\Phi^{-1}(\mu)$ where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable), or complementary log-log (i.e. $\log(-\log(1 - \mu))$)

link function for the generalized linear models (McCullagh, 2019) since these link functions map the interval $(0, 1)$ onto the whole real line $(-\infty, +\infty)$. When the probability $P(Y_{ik} = 1 \mid \mathbf{Z}_{ik}; \boldsymbol{\beta})$ is very small for the whole range of $\mathbf{Z}_{ik}$ the natural log link function can be considered under the Poisson approximation. Hartmann et al (2021) report rates of significant hypotensive adverse events of 0.051% (control) and 0.035% (experimental) for this setting. With the response probability being low we specify regression models of the form

$$\log(\mu(\mathbf{Z}_{ik}; \boldsymbol{\beta})) = \beta_0 + \beta_1 X_i + \boldsymbol{\beta}'_2 \mathbf{V}_{ik} = \boldsymbol{\beta}' \mathbf{Z}_{ik}, \quad (3.1)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \boldsymbol{\beta}'_2)'$. Under this link the estimand

$$\beta_1 = \log \frac{P(Y_{ik} = 1 \mid X_i = 1, \mathbf{V}_{ik}; \boldsymbol{\beta})}{P(Y_{ik} = 1 \mid X_i = 0, \mathbf{V}_{ik}; \boldsymbol{\beta})}$$

is the log relative rate of an adverse event for the experimental arm compared to the control arm, conditional on the covariate $\mathbf{V}_{ik}$. This link function yields a model with the appealing property of being collapsible (Geng et al, 2001), meaning that the causal effect β_1 has the same interpretation in nested models obtained by removing covariates in $\mathbf{V}_{ik}$ provided $X_i \perp \mathbf{V}_{ik}$, which is the case here due to randomization.

The generalized estimating equations (GEE) (Liang and Zeger, 1986) for $\boldsymbol{\beta}$ can be written as

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{U}_i(\mathbf{Z}_i; \boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{D}'(\mathbf{Z}_i; \boldsymbol{\beta}) \boldsymbol{\Sigma}(\mathbf{Z}_i; \boldsymbol{\beta})^{-1} \{\mathbf{Y}_i - \boldsymbol{\mu}(\mathbf{Z}_i; \boldsymbol{\beta})\} = 0, \quad (3.2)$$

where $\boldsymbol{\mu}(\mathbf{Z}_i; \boldsymbol{\beta}) = (\mu(\mathbf{Z}_{i1}; \boldsymbol{\beta}), \dots, \mu(\mathbf{Z}_{iK_i}; \boldsymbol{\beta}))'$ is a $K_i \times 1$ vector of means indexed by parameter $\boldsymbol{\beta}$ for the effects of covariates, $\mathbf{D}(\mathbf{Z}_i; \boldsymbol{\beta}) = \partial \boldsymbol{\mu}(\mathbf{Z}_i; \boldsymbol{\beta}) / \partial \boldsymbol{\beta}'$ is a $K_i \times p$ matrix, and $\boldsymbol{\Sigma}(\mathbf{Z}_i; \boldsymbol{\beta}) = \text{cov}(\mathbf{Y}_i \mid \mathbf{Z}_i)$ is the $K_i \times K_i$ covariate matrix of the responses from individual i , $i = 1, \dots, n$. Under the working independence assumption with the Poisson (identity) variance function the working covariance matrix $\boldsymbol{\Sigma}(\mathbf{Z}_i; \boldsymbol{\beta})$ is diagonal and indexed by $\boldsymbol{\beta}$. Note that in practice the standard error of $\hat{\boldsymbol{\beta}}$ is obtained from the estimated robust covariance matrix for $\hat{\boldsymbol{\beta}}$ given by

$$\widehat{\text{cov}}(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) = \hat{\mathbf{A}}^{-1}(\boldsymbol{\beta}) \hat{\mathbf{B}}(\boldsymbol{\beta}) \{\hat{\mathbf{A}}^{-1}(\boldsymbol{\beta})\}' \quad (3.3)$$

where

$$\hat{\mathbf{A}}(\boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\beta}'} \mathbf{U}_i(\mathbf{Z}_i; \boldsymbol{\beta}) \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}},$$

and

$$\widehat{B}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{U}_i(\mathbf{Z}_i; \boldsymbol{\beta}) \mathbf{U}_i'(\mathbf{Z}_i; \boldsymbol{\beta}) \Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}.$$

Our interest lies primarily in $\widehat{\beta}_1$, the estimated coefficient of X_i , since $\exp(\widehat{\beta}_1)$ is an estimate of the relative rate of an adverse event in the experimental to control arms. The standard error of $\widehat{\beta}_1$ is obtained from the square root of the corresponding entry of (3.3) and based on this valid tests of $H_0 : \beta_1 = 0$ can be carried out and confidence intervals for β_1 can be obtained with good empirical coverage.

3.2.2 Large Sample Results for Study Design

At the design stage of a study, we aim to ensure power objectives are met under full model assumptions. In particular it is necessary to design trials in anticipation of the sampling variation for the primary parameter estimator at study completion. Here this is made explicit by the need to evaluate the asymptotic covariance matrix in (3.4),

$$\sqrt{n} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) \xrightarrow{d} \text{MVN} \left(0, \mathcal{A}^{-1}(\boldsymbol{\beta}) \mathcal{B}(\boldsymbol{\beta}) \{ \mathcal{A}^{-1}(\boldsymbol{\beta}) \}' \right) \quad \text{as } n \rightarrow \infty, \quad (3.4)$$

where

$$\mathcal{A}(\boldsymbol{\beta}) = -\mathbb{E} \left\{ \frac{\partial}{\partial \boldsymbol{\beta}'} \mathbf{U}_i(\mathbf{Z}_i; \boldsymbol{\beta}) \right\} = \mathbb{E} \left\{ D_i'(\mathbf{Z}_i; \boldsymbol{\beta}) \Sigma_i^{-1}(\mathbf{Z}_i; \boldsymbol{\beta}) D(\mathbf{Z}_i; \boldsymbol{\beta}) \right\}, \quad (3.5)$$

and

$$\mathcal{B}(\boldsymbol{\beta}) = \text{cov}(\mathbf{U}_i(\mathbf{Z}_i; \boldsymbol{\beta}), \mathbf{U}_i(\mathbf{Z}_i; \boldsymbol{\beta})) = \mathbb{E} \left\{ \mathbf{U}_i(\mathbf{Z}_i; \boldsymbol{\beta}) \mathbf{U}_i'(\mathbf{Z}_i; \boldsymbol{\beta}) \right\}, \quad (3.6)$$

are $p \times p$ matrices. The expectations in (3.5) and (3.6) are taken with respect to the true data generating model which must be specified at the design stage in a way that recognizes all sources of variability in the data. The matrices $\mathcal{A}(\boldsymbol{\beta})$ and $\mathcal{B}(\boldsymbol{\beta})$ can be seen to depend on $\boldsymbol{\beta}$, the association between the responses conditional on $\mathbf{Z}_i$, the treatment allocation and distribution of the prognostic variables $\mathbf{Z}_i$, along with the distribution of K_i . Note that the distribution of X_i is set by the allocation ratio (i.e. the randomization ratio), and $X_i \perp \mathbf{Z}_i$ due to randomization, but we typically aim to avoid making assumptions about covariate distributions or their effect on the outcome when planning clinical trials. In the next section we consider the calculations and design for a simplified setting in which only X_i is included in the regression model for the outcome - in practice the robust variance estimate in (3.3) will ensure valid inferences even when important prognostic factors are

omitted.

3.3 A Simplified Setting for Trial Design

3.3.1 Large Sample Results

In clinical trials we often consider a simplified model at the design stage in part for robustness and to reduce the number of processes that must be modeled. To avoid having to specify a distribution for $\mathbf{V}_{ik}$ here we adopted a simplified version of (3.1) with $\beta_2 = \mathbf{0}$. In this case the means of responses over repeated measurements are the same (i.e. $\mu(\mathbf{Z}_{ik}; \beta) = \mu(X_i; \beta)$). We can then replace $\mathbf{Z}_{ik}$ and $\mathbf{Z}_i$ by X_i in the following and let $\beta = (\beta_0, \beta_1)'$. We then have

$$D(X_i; \beta) = \frac{\mu(X_i; \beta)}{\partial \beta'} = \mu(X_i; \beta) \begin{pmatrix} 1 & X_i \\ \vdots & \vdots \\ 1 & X_i \end{pmatrix}.$$

With the Poisson (identity) variance function and working independence assumptions the $K_i \times K_i$ working covariance matrix has identical entries in the diagonal and is given by

$$\Sigma(X_i; \beta) = \begin{pmatrix} \mu(X_i; \beta) & & 0 \\ & \ddots & \\ 0 & & \mu(X_i; \beta) \end{pmatrix}.$$

The generalized estimating function in (3.2) can then be simplified to

$$\mathbf{U}(\beta) = \sum_{i=1}^n \mathbf{U}_i(X_i; \beta) = \sum_{i=1}^n \sum_{j=1}^{K_i} \begin{pmatrix} U_{ij1}(\beta) \\ U_{ij2}(\beta) \end{pmatrix},$$

where $U_{ij1}(\beta) = Y_{ij} - \mu(X_i; \beta)$ and $U_{ij2}(\beta) = X_i(Y_{ij} - \mu(X_i; \beta))$.

For variance estimation expectations are required with respect to all elements of the data. In what follows we consider the evaluation of (3.5) and (3.6) for planning a study with these estimating functions defining the estimator at the end of the study. Regarding the number of donations we let $\kappa_x = \mathbb{E}(K_i \mid X_i = x)$ be the expected number of donations of an individual in group x , and $\nu_x = \text{var}(K_i \mid X_i = x)$ be the variance of the number of

donations for a subject in group x , $x = 0, 1$. For the response we let $\mu_x = \mu(X_i = x; \beta)$ and let $V_x = \text{var}(Y_{ij} \mid X_i = x)$ be the true variance of responses for the subject in group x so that $V_x = \mu_x(1 - \mu_x)$, $x = 0, 1$. Furthermore if we assume an exchangeable pairwise correlation for responses within individuals (i.e., $\text{corr}(Y_{ij}, Y_{ik} \mid X_i = x) = \rho$) then $\text{cov}(Y_{ij}, Y_{ik} \mid X_i = x) = \rho V_x$.

In Appendix B.1 we show that the $\mathcal{A}(\beta)$ matrix in (3.5) is

$$\mathcal{A}(\beta) = \begin{pmatrix} \sum_{x=0}^1 \kappa_x \mu_x P(X_i = x) & \kappa_1 \mu_1 P(X_i = 1) \\ \kappa_1 \mu_1 P(X_i = 1) & \kappa_1 \mu_1 P(X_i = 1) \end{pmatrix},$$

and $\mathcal{B}(\beta)$ in (3.6) is

$$\begin{pmatrix} \sum_{x=0}^1 \{ \kappa_x V_x + (\nu_x + \kappa_x^2 - \kappa_x) \rho V_x \} P(X_i = x) & \{ \kappa_1 V_1 + (\nu_1 + \kappa_1^2 - \kappa_1) \rho V_1 \} P(X_i = 1) \\ \{ \kappa_1 V_1 + (\nu_1 + \kappa_1^2 - \kappa_1) \rho V_1 \} P(X_i = 1) & \{ \kappa_1 V_1 + (\nu_1 + \kappa_1^2 - \kappa_1) \rho V_1 \} P(X_i = 1) \end{pmatrix}.$$

Then $\mathcal{A}^{-1}(\beta) \mathcal{B}(\beta) \{\mathcal{A}^{-1}(\beta)\}'$ in (3.4) can be written as

$$\mathcal{A}^{-1}(\beta) \mathcal{B}(\beta) \{\mathcal{A}^{-1}(\beta)\}' = \begin{pmatrix} \frac{V_0[\kappa_0 + (\nu_0 + \kappa_0^2 - \kappa_0)\rho]}{\kappa_0^2 \mu_0^2 P(X_i = 0)} & -\frac{V_0[\kappa_0 + (\nu_0 + \kappa_0^2 - \kappa_0)\rho]}{\kappa_0^2 \mu_0^2 P(X_i = 0)} \\ -\frac{V_0[\kappa_0 + (\nu_0 + \kappa_0^2 - \kappa_0)\rho]}{\kappa_0^2 \mu_0^2 P(X_i = 0)} & \sum_{x=0}^1 \frac{V_x[\kappa_x + (\nu_x + \kappa_x^2 - \kappa_x)\rho]}{\kappa_x^2 \mu_x^2 P(X_i = x)} \end{pmatrix}. \quad (3.7)$$

The asymptotic variance of $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ is the (2, 2) entry of (3.7) which is

$$\Omega_{\beta_1 \beta_1} = \frac{V_0}{\kappa_0 \mu_0^2 P(X_i = 0)} \left\{ 1 + \frac{(\nu_0 + \kappa_0^2 - \kappa_0)\rho}{\kappa_0} \right\} + \frac{V_1}{\kappa_1 \mu_1^2 P(X_i = 1)} \left\{ 1 + \frac{(\nu_1 + \kappa_1^2 - \kappa_1)\rho}{\kappa_1} \right\}. \quad (3.8)$$

Note that the marginal variance of each binomial response is $V_x = \mu_x(1 - \mu_x)$, $x = 0, 1$, so (3.8) can be further simplified to

$$\Omega_{\beta_1 \beta_1} = \frac{1 - \mu_0}{\kappa_0 \mu_0 P(X_i = 0)} \left\{ 1 + \frac{(\nu_0 + \kappa_0^2 - \kappa_0)\rho}{\kappa_0} \right\} + \frac{1 - \mu_1}{\kappa_1 \mu_1 P(X_i = 1)} \left\{ 1 + \frac{(\nu_1 + \kappa_1^2 - \kappa_1)\rho}{\kappa_1} \right\}. \quad (3.9)$$

Since $\nu_x + \kappa_x^2 - \kappa_x > 0$, a correlation $\rho > 0$ will inflate the variance in (3.9). Moreover (3.9)

is decreasing in κ_0 and κ_1 so the more donations provided by each recruited individual, the more precise the estimate of β_1 . If the responses within the individual are independent, balanced randomization is used, and the mean number of donations per recruited individual is the same in the two arms ($\kappa_0 = \kappa_1 = \kappa$, say), then (3.9) reduces to

$$\frac{2}{\kappa} \left(\frac{1 - \mu_0}{\mu_0} + \frac{1 - \mu_1}{\mu_1} \right).$$

The general form of (3.9) is best to use for the purpose of trial planning.

3.3.2 Zero-Truncated Conditional Negative Binomial Process for Donations

As shown in (3.8) the mean and variance of the number of measurements per individual are required for the evaluation of the asymptotic variance which is used for the sample size calculation. To develop this we consider donations as arising in the population of donors according to a mixed Poisson process (Cook and Lawless, 2007) in which each donor has their own intensity (rate), $U\lambda$, for donation, where λ is the population mean rate, and $U > 0$ is a random effect with $\mathbb{E}(U) = 1$ and $\text{var}(U) = \phi > 0$. If K is the number of donations for an individual over a period of duration τ then $P(K = k | u) = (u\lambda\tau)^k e^{-u\lambda\tau} / k!$, $k = 0, 1, \dots$. If U is gamma distributed with density function $g(u; \phi) = u^{\phi^{-1}-1} \exp(-u/\phi) / (\Gamma(\phi^{-1})\phi^{\phi^{-1}})$, marginalizing over U gives the negative binomial probability mass function (Lawless, 1987)

$$P(K = k; \lambda, \phi) = \frac{\Gamma(\phi^{-1} + k)}{\Gamma(\phi^{-1}) k!} \left(\frac{\lambda\tau\phi}{1 + \lambda\tau\phi} \right)^k \left(\frac{1}{1 + \lambda\tau\phi} \right)^{\phi^{-1}}, \quad (3.10)$$

with mean $\mathbb{E}(K) = \lambda\tau$ and variance $\text{var}(K) = \lambda\tau(1 + \lambda\tau\phi)$.

Figure 3.1 depicts the recruitment and follow-up phases of a trial with the recruitment period of length τ_A and a follow-up period of length τ_F for each recruited individual. The horizontal axis records the time since the start of the study (in calendar time) and the vertical axis records the time on study for recruited individuals. The 45° lines emanating from the x-axis represent the passage of calendar time and study follow-up time with the points of intersection with x-axis corresponding to the date individuals were encountered upon their initial donation. The dots on the 45° lines represent donation occasions during the follow-up phase, and the response of interest (Y_{ik}) is recorded at the k th donation which occurs at such times. Let K_A denote the number of donation occasions from a random donor in the population which occur during the recruitment phase of length τ_A , and K_F

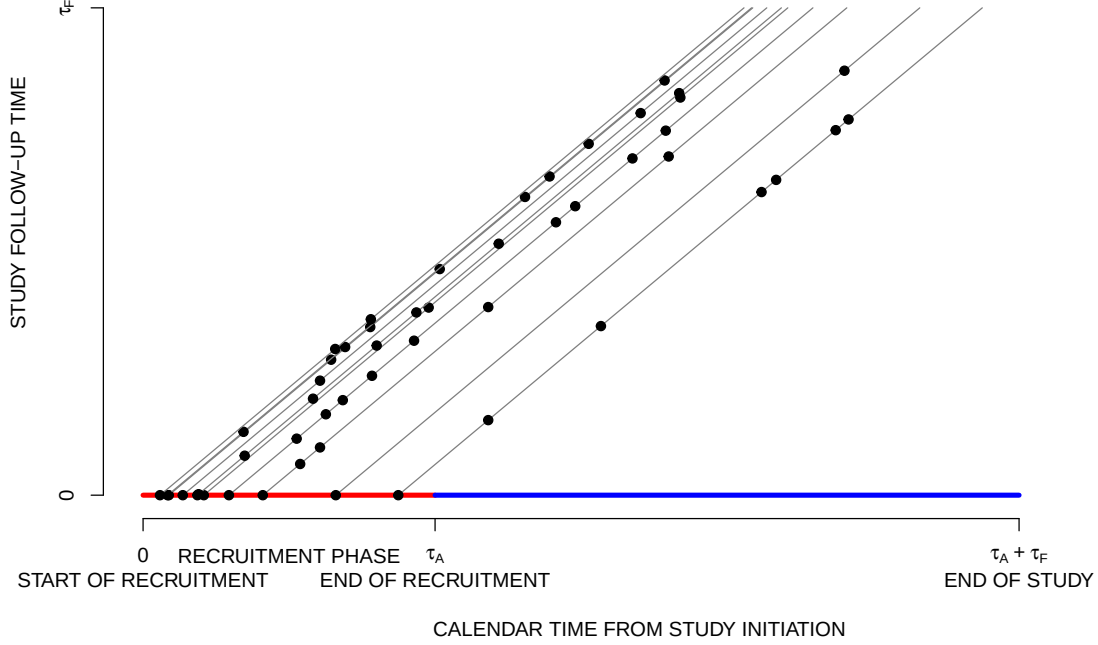


Figure 3.1: An illustration of 10 individuals recruitment process and repeated measurements.

be the number of donations from the subject in the follow-up phase of length τ_F . Such counts will be positively correlated since each individual will have their common random effect affecting their donation rate in the two time periods ([Cook and Wei, 2002](#)).

To be recruited to a study, however, individuals must have at least one donation during the recruitment period, so the number of donations in the recruitment phase among recruited individuals is zero-truncated. Our interest is the mean and variance of the number of donations per subject over the follow-up period among subjects encountered (i.e. having at least one donation) and recruited during the recruitment phase. Since we require $K_A \geq 1$ to be included in the study, we are interested in the conditional mean and variance of the number of donations K_F over the follow-up period of duration τ_F . With the mixed Poisson model, $K_A | u \sim \text{Poisson}(u\lambda\tau_A)$ and $K_F | u \sim \text{Poisson}(u\lambda\tau_F)$ and the model for the number of donations during the study can be shown to have a probability

mass function $P(K_F = k \mid K_A \geq 1; \lambda, \phi)$ of the form

$$\frac{\Gamma(\phi^{-1} + k)}{\Gamma(\phi^{-1})} \frac{1}{k!} \frac{1}{(1 + \lambda\tau_A\phi)^{\phi^{-1}} - 1} \left\{ \left(\frac{\lambda\tau_F\phi}{1 + \lambda\tau_F\phi} \right)^k \left(\frac{1 + \lambda\tau_A\phi}{1 + \lambda\tau_F\phi} \right)^{\phi^{-1}} \right. \\ \left. \left(\frac{\lambda\tau_F\phi}{1 + \lambda\tau_F\phi + \lambda\tau_A\phi} \right)^k \left(\frac{1 + \lambda\tau_A\phi}{1 + \lambda\tau_F\phi + \lambda\tau_A\phi} \right)^{\phi^{-1}} \right\}, \quad (3.11)$$

where k denotes the number of donations during the follow-up phase. Individual-level data on donation frequencies (number of donations/month for established donors) are available for the population of donors from administrative records maintained by independent donation centers and blood suppliers such as Canadian Blood Services. With such data over a period of calendar time of length τ , λ and ϕ can be estimated by maximizing a likelihood based on (3.11).

For each value of λ and ϕ , and duration of recruitment and follow-up periods τ_A and τ_F , the mean of K_F given $K_A \geq 1$ can be calculated based on (3.11) as

$$\mathbb{E}(K_F \mid K_A \geq 1) = \lambda\tau_F \left\{ \frac{1 - (1 + \lambda\tau_A\phi)^{-\phi^{-1}-1}}{1 - (1 + \lambda\tau_A\phi)^{-\phi^{-1}}} \right\}, \quad (3.12)$$

and the variance of K_F given $K_A \geq 1$ is

$$\nu = \text{var}(K_F \mid K_A \geq 1) = \mathbb{E}(K_F \mid K_A \geq 1) - \mathbb{E}^2(K_F \mid K_A \geq 1) + \\ (\lambda\tau_F)^2(1 + \phi) \left\{ \frac{1 - (1 + \lambda\tau_A\phi)^{-\phi^{-1}-2}}{1 - (1 + \lambda\tau_A\phi)^{-\phi^{-1}}} \right\}. \quad (3.13)$$

The donation during the recruitment period is counted as a study donation, so we define $K = 1 + K_F$ giving $\kappa = 1 + \mathbb{E}(K_F \mid K_A \geq 1)$. The details of these derivations are provided in Appendix B.2.

We let L denote the time that an individual arrives at their donation center after the start of the study. Based on the mixed Poisson formulation of the donation process we have $L \mid U \sim \exp(U\lambda)$ and marginalizing over U gives the density of a Pareto distribution

(Lawless, 2011) for the accrual time

$$f(l; \lambda, \phi) = \lambda \left(\frac{1}{\lambda l \phi + 1} \right)^{\phi^{-1}+1}.$$

The requirement that a donor must arrive to their donation center to provide at least one donation during the recruitment period in order to be recruited means $L \leq \tau_A$ for recruited individuals, and

$$P(L \leq \tau_A) = 1 - (\lambda \phi \tau_A + 1)^{-\phi^{-1}}.$$

This means that $\mathbb{E}(K_F \mid K_A \geq 1) = \mathbb{E}(K_F \mid L \leq \tau_A)$ and $\text{var}(K_F \mid K_A \geq 1) = \text{var}(K_F \mid L \leq \tau_A)$. In Appendix B.3 we provide the details of related derivations.

For planning purposes it is helpful to consider the distribution of the sample size as a function of time since study initiation. The expected number of donors encountered by time τ since the start of the accrual period, denoted by $N_A(\tau)$, is a function of the size of the study donor population N_{pop} and the parameters characterizing the population donation process. It is given by

$$\mathbb{E}\{N_A(\tau)\} = N_{\text{pop}} \cdot P(L \leq \tau) = N_{\text{pop}} \cdot \{1 - (\lambda \phi \tau + 1)^{-\phi^{-1}}\}, \quad (3.14)$$

which for a specified N_{pop} , λ and ϕ , it is an increasing function of τ converging to N_{pop} as $\tau \rightarrow \infty$. To accommodate the fact that not all encountered donors will consent to participate in the study we let η represent the consent rate and define the expected number of recruited donors over $(0, \tau]$ to be

$$\mathbb{E}\{N(\tau)\} = \mathbb{E}\{N_A(\tau)\}\eta.$$

We can then find the value of τ_A by which time a sufficient sample can be expected so that the power requirements are achieved. Note that in this framework the study aims to find the duration of the accrual period and the sample size of donors recruited over this period is random. One can alternatively set the sample size equal to the expected number of recruited individuals and let the accrual period end when this sample size is achieved; in the latter case the duration of the accrual period is random. We consider both implementations of this design in Section 3.4.

We stress here that for a given setting defined by the study donor population size N_{pop} , λ and ϕ , the design parameters τ_A and τ_F determine both the types of donors recruited and the average number of donations that recruited donors provide during study follow-up. So if τ_A is small and τ_F is large, then “high frequency” donors are preferentially recruited

and such individuals will be expected to provide responses for a large number of donations over the study period. If τ_F is smaller, then the number of donors will necessarily have to increase, and τ_A will have to be long enough to ensure that enough individuals can be recruited to satisfy power objectives. Note that in what follows we consider the donor population size N_{pop} as fixed; in practice this may be set by the number of donation sites enlisted to take part in the study. Trials may be conducted very rapidly if many plasma donation sites are enlisted; all things being equal, doubling the size of the donor pool by increasing the number of sites to $2N_{\text{pop}}$ will double the accrual rate and enable a shortening of the accrual period. We consider the sample size as a random variable $N(\tau)$ here, but in the simulation studies that follow we show that one can equivalently consider designs with samples fixed at $n = \mathbb{E}\{N(\tau)\}$, the expected number of donors by the end of the derived tentative accrual period; then one can define the end of the accrual period adaptively based on when the n th individual is recruited. In this case the mean accrual period is extremely close to the calculated τ_A provided the follow-up period remains at τ_F .

3.4 Study Design, Validation and an Application

3.4.1 Sample Size for Superiority and Non-inferiority Trials

Given the donor population size N_{pop} , follow-up duration τ_F , and consent rate η , the expected sample size $\mathbb{E}\{N(\tau)\}$ can be derived as a function of the duration of the recruitment period τ_A using (3.14).

For a superiority trial we consider a two-sided test with the aim to test

$$H_0 : \beta_1 = \beta_1^\circ \text{ vs. } H_A : \beta_1 \neq \beta_1^\circ .$$

With the significance level α , the desired power can be written as

$$\begin{aligned} \text{Power} &= P \left(\left| \frac{\hat{\beta}_1 - \beta_0}{\sqrt{\Omega_{\beta_{1A}\beta_{1A}}/\mathbb{E}\{N(\tau_A)\}}} \right| > z_{1-\alpha/2} \mid \beta_1 = \beta_{1A} \right) \\ &= 1 - \Phi \left(z_{1-\alpha/2} - \frac{\beta_{1A}}{\sqrt{\Omega_{\beta_{1A}\beta_{1A}}/\mathbb{E}\{N(\tau_A)\}}} \right) + \Phi \left(-z_{1-\alpha/2} - \frac{\beta_{1A}}{\sqrt{\Omega_{\beta_{1A}\beta_{1A}}/\mathbb{E}\{N(\tau_A)\}}} \right), \end{aligned} \quad (3.15)$$

where $\Omega_{\beta_{1A}\beta_{1A}}$ is obtained from (3.8) under the alternative hypothesis, β_{1A} is the true effect, $z_{1-\alpha/2}$ is the $100(1 - \alpha/2)\%$ percentile of the standard normal distribution, and $\Phi(\cdot)$ is the

cumulative distribution function of the standard normal random variable. Here we consider the study population donation process both to learn about the rate of recruitment to plan the duration of the accrual period τ_A and also to infer what the mean and variance will be for the number of donations per recruited donor. This enables specification of the τ_A value which will ensure the design satisfies the power objectives. In this case the duration of the accrual period is fixed but when we consider the conduct of a study the actual number of donors will be random. If it is preferable to plan a study based on sample size then one can fix the target sample size as the mean $\mathbb{E}\{N(\tau_A)\}$ in which case the accrual period can terminate when the n th donor is recruited; this may lead to some studies having a shorter and some studies having a longer accrual period. We refer to the former implementation as designs having a fixed τ_A and the latter as designs with a fixed sample size n where $n = \mathbb{E}\{N(\tau_A)\}$ and we investigate both in the simulations that follow.

A non-inferiority design is considered when the goal is to assess the efficacy of an experimental treatment that is expected to be not unacceptably worse to the standard treatment, or when it is unethical to conduct a placebo-controlled trial. Unlike the superiority design where the null hypothesis states no difference between treatments, the null hypothesis of a non-inferiority trial states that the experimental treatment is inferior to the standard treatment by more than a specified non-inferiority margin (James Hung et al, 2003). If the null hypothesis is rejected then it can be concluded that the experimental treatment is non-inferior to the standard treatment. We consider a margin β_{1M} which is the log-relative rate at the margin value of β_1 , i.e. $\beta_{1M} = \log RR_M$, where for $RR < RR_M$ we can claim non-inferiority on the relative rate scale. This corresponds to the case where the relative rate of an adverse event for the blood donation using the experimental system versus the standard system is no greater than RR_M . For a non-inferiority design we consider a one-sided test with the aim to test

$$H_0 : \beta_1 \geq \beta_{1M} \text{ vs. } H_A : \beta_1 < \beta_{1M}.$$

Here the sample size n can be determined to ensure adequate power where

$$\begin{aligned} \text{Power} &= P \left(\frac{\hat{\beta}_1 - \beta_{1M}}{\sqrt{\Omega_{\beta_{1A}\beta_{1A}}/\mathbb{E}\{N(\tau_A)\}}} < -z_{1-\alpha/2} \mid \beta_1 = \beta_{1A} \right) \\ &= \Phi \left(\frac{\beta_{1M} - \beta_{1A}}{\sqrt{\Omega_{\beta_{1A}\beta_{1A}}/\mathbb{E}\{N(\tau_A)\}}} - z_{1-\alpha/2} \right), \end{aligned} \quad (3.16)$$

and $\beta_{1A} < \beta_{1M}$ is the true effect value within the non-inferiority region. Again non-

inferiority studies can be designed with fixed accrual period (and random sample size) or with the fixed sample size set at the expected number of recruited donors at expected τ_A . We investigate both implementations in what follows.

3.4.2 Empirical Studies

Here we report on the design and findings of simulation studies geared towards validation of the guidelines for study planning. We first tabulate the calculated durations of the accrual period and expected sample sizes as a function for a number of different scenarios. We consider donor populations of size $N_{\text{pop}} = 40,000$ or $60,000$. We suppose the recruitment rate is 100% (i.e. $\eta = 1$) and recruited donors are assigned through balanced randomization so $P(X_i = 1) = P(X_i = 0) = 0.5$. We suppose time is measured in years and let $\tau_F = 0.5$ denote the duration of the follow-up period. We consider annual donation rates of $\lambda = 10$ or 20 with moderate ($\phi = 0.8$) and stronger ($\phi = 1.6$) degrees of heterogeneity in donation rates in the population. Serious hypotensive adverse events are very rare and we set the control event rate to $\mu_0 = 0.002$ and the serial correlation to $\rho = 0.005$ or 0.01 . For the case where a superiority trial is planned we set $\beta_1^o = 0$. A more common setting is a non-inferiority trial in which case we set the margin to $\beta_{1M} = \log(2)$ so that the experimental algorithm is considered non-inferior to the standard algorithm as long as it no more than doubles the adverse event rate; this is a large margin in a relative sense, but with the event rate so low it corresponds to an absolute increase of only 0.2%. An elevation in the adverse event rate may be expected and so we set $\beta_{1A} = \log(1.5)$ to cover this possibility and design the trials to ensure adequate power at the 80% or 90% levels when the type I error rate is controlled at $\alpha = 5\%$.

We first derive the appropriate duration of the accrual period based on the specified parameters using the large sample variance given in (3.8) and solving equations in (3.15) and (3.16) under the specified scenarios. The calculated necessary recruitment duration (τ_A) and expected sample size ($\mathbb{E}\{N(\tau_A)\}$) are given in columns 8 and 9 of Table 3.1 for a superiority trial with target powers of 80% and 90% respectively. In the preceding columns 6 and 7 are the expected number of donations and the standard deviations for the number of donations per donor respectively. As expected we see that the accrual phase is shorter with larger donor populations and when the mean annual donation rate is greater, the number of donors required is lower. Interestingly we see that the greater the variation in the annual donation rate in the population the shorter the recruitment period and the fewer donations are required. This arises since donors with a higher individual-specific donation rate (i.e. a larger random effect) will tend to be recruited early in the accrual period in this case, and such individuals will give correspondingly more donations (on average) in

the follow-up period. The corresponding designs are given for non-inferiority testing in Table 3.2.

To validate the calculations we next assess the empirical rejection rates to verify the design calculations. Under each parameter setting, we consider two implementations of the trial design. We consider the design where the sample size is fixed at $n = \mathbb{E}\{N(\tau_A)\}$ (the “fixed n ” design) whereby we generate a random effect for each individual in the donor population as $U_i \sim \text{Gamma}(\phi^{-1}, \phi)$ and a recruitment time as $l_i \sim \exp(u_i \lambda)$ for $i = 1, \dots, N_{\text{pop}}$. The recruitment times are then sorted in ascending order and we retain the first n donors for the study assuming a 100% consent rate for illustration (i.e. $\eta = 1$). The recruitment duration is then determined adaptively by the recruitment time of the n th individual. Thus in this case we treat the sample size n as fixed and the recruitment duration as random. The alternative approach is to consider the accrual period as fixed at the derived τ_A and to consider the number of recruited individuals over $(0, \tau_A]$ as random – this is in line with the nature of the calculations; we refer to this as the “fixed τ_A ” design. In this implementation we only retain the individuals whose recruitment times fall before the derived end of the recruitment period (τ_A). The following simulations confirm that these two implementations perform equally well.

For either method, upon recruitment each individual i in the sample is assigned to treatment via $X_i \sim \text{Bernoulli}(0.5)$. The number of donations during the follow-up period for each subject is generated by $K_{iF} \sim \text{Poisson}(u_i \lambda \tau_F)$. Thus the number of donations for subject i in the study is $K_{iF} + 1$ as we assume subjects makes their first donation upon entering the study. We adopt a joint model for the binary responses $\{Y_{ik}, k = 1, 2, \dots, K_i\}$ to accommodate the serial dependence within individuals (Qaqish, 2003; Preisser and Qaqish, 2014). Specification of the joint probability model for $\mathbf{Y}_i \mid \mathbf{Z}_i$ involves specification of a marginal mean conditional on covariates, along with the conditional association characterizing the serial dependence. For a randomly generated Y_{i1} from Bernoulli distribution with mean $P(Y_{ik} = 1 \mid \mathbf{Z}_{i1})$, the response Y_{ik} , $k = 2, 3, \dots, K_i$, is generated sequentially as Bernoulli with the conditional mean given by Qaqish (2003)

$$\mathbb{E}(Y_{ik} \mid \bar{\mathbf{Y}}_{ik}, \bar{\mathbf{Z}}_{ik}, \mathbf{Z}_{ik}) = \mathbb{E}(Y_{ik} \mid \mathbf{Z}_{ik}) + \sum_{j=1}^{k-1} b_{ikj}(y_{ij} - \mathbb{E}(Y_{ij} \mid \mathbf{Z}_{ij})), \quad k = 2, 3, \dots, K_i,$$

where $\bar{\mathbf{Y}}_{ik} = (Y_{i1}, \dots, Y_{i,k-1})'$, $\bar{\mathbf{Z}}_{ik} = (\mathbf{Z}_{i1}, \dots, \mathbf{Z}_{i,k-1})'$, and b_{ikj} is the j th element of

$$\mathbf{b}_{ik}(\bar{\mathbf{Z}}_{ik}, \mathbf{Z}_{ik}) = \text{cov}(\bar{\mathbf{Y}}_{ik} \mid \bar{\mathbf{Z}}_{ik})^{-1} \text{cov}(\bar{\mathbf{Y}}_{ik}, Y_{ik} \mid \bar{\mathbf{Z}}_{ik}, \mathbf{Z}_{ik}), \quad k = 2, 3, \dots, K_i.$$

Thus under our parameter settings the correlated binary responses for individual i can

Table 3.1: Length of recruitment duration (τ_A) and expect number of donors ($\mathbb{E}(N)$) required across sites for a superiority trial with $\mu_0 = 0.002$, $\exp(\beta_{1A}) = 1.5$, $\tau_F = 0.5$, and $\beta_1^\circ = 0$; verification by calculation of the empirical type I error and the empirical power; $n_{sim} = 1000$.

Parameter setting					Analytical results				Empirical results (fixed n) [†]			Empirical results (fixed τ_A) [†]			
N_{pop}	Target power (%)	λ	ϕ	ρ	κ	ν	τ_A	$\mathbb{E}\{N(\tau_A)\}$	$\hat{\tau}_A$	Emp. power (%)	Emp. type I error (%)	$\hat{\mathbb{E}}\{N(\tau_A)\}$	Emp. power (%)	Emp. type I error (%)	
40,000	80	10	0.80	0.005	9.16	38.09	0.029	9203	0.029	80.7	4.5	9203	80.0	4.7	
				0.010	9.10	37.68	0.031	9786	0.031	80.9	3.8	9784	81.4	4.9	
				1.60	0.005	12.22	90.35	0.023	7097	0.023	80.8	5.4	7100	80.8	4.4
		20	0.80	0.010	12.05	88.18	0.026	7802	0.026	81.5	5.7	7795	78.2	4.1	
				0.005	18.10	147.49	0.007	4945	0.007	81.4	5.3	4946	82.8	6.0	
				0.010	18.00	145.89	0.008	5518	0.008	80.0	5.1	5518	82.2	3.1	
	90	10	0.80	0.005	25.07	382.51	0.005	3789	0.005	82.3	4.9	3788	79.0	6.2	
				0.010	24.73	373.08	0.007	4459	0.007	80.7	5.2	4460	78.5	5.6	
				0.010	8.83	35.72	0.045	12754	0.045	92.1	4.0	12748	91.6	4.6	
		20	0.80	0.010	8.75	35.20	0.049	13580	0.049	90.6	5.2	13578	92.7	5.1	
				0.005	11.53	81.83	0.037	10031	0.037	90.6	6.8	10034	91.3	4.1	
				0.010	11.29	79.20	0.042	11039	0.042	91.1	5.1	11040	92.5	4.5	
	60,000	80	10	0.80	0.005	9.46	40.40	0.017	8929	0.017	80.7	4.8	8924	80.8	4.9
					0.010	9.42	40.13	0.018	9482	0.018	79.6	5.3	9486	81.3	4.8
					1.60	0.005	12.85	98.97	0.013	6786	0.013	81.9	5.4	6789	82.3
			20	0.80	0.010	12.75	97.46	0.015	7427	0.015	82.0	3.6	7429	79.3	4.2
					0.005	18.41	152.32	0.004	4874	0.004	79.7	6.7	4872	82.6	5.4
					0.010	18.35	151.28	0.005	5415	0.005	82.8	4.1	5416	78.8	5.3
		90	10	0.80	0.005	25.72	401.66	0.003	3738	0.003	81.8	5.3	3739	81.2	4.3
					0.010	25.52	395.71	0.004	4329	0.004	80.6	5.7	4326	79.1	4.7
					0.005	9.26	38.83	0.025	12195	0.025	91.6	5.2	12203	90.3	3.9
			20	0.80	0.010	9.21	38.47	0.027	12975	0.027	91.0	5.9	12975	89.0	5.0
					0.005	12.43	93.09	0.019	9364	0.019	92.0	5.9	9366	89.8	5.3
					0.010	12.28	91.14	0.022	10269	0.022	90.4	6.3	10266	89.8	5.3
90		20	0.80	0.005	18.21	149.09	0.006	6565	0.006	91.2	6.1	6562	90.2	5.3	
				0.010	18.11	147.60	0.007	7357	0.007	90.2	4.6	7357	90.7	5.4	
				1.60	0.005	25.27	388.45	0.005	5066	0.005	91.0	3.9	5068	91.4	4.7
		0.010	24.99	380.28	0.006	5917	0.006	90.1	3.9	5914	90.1	4.9			

[†] Columns headed “fixed n ” recruited donors until the expected number of donors $\mathbb{E}\{N(\tau_A)\}$ were recruited and at that point the recruitment period ended (and so τ_A is random). The average duration of the random recruitment period $\hat{\tau}_A$ was very close to τ_A . For columns headed “fixed τ_A ” the accrual period was fixed at the derived τ_A and the actual number of donors recruited varies across the $n_{sim} = 1000$ simulations for each row of the table.

Table 3.2: Length of recruitment duration (τ_A) and expect number of donors ($\mathbb{E}(N)$) required across sites for a non-inferiority trial with $\mu_0 = 0.002$, $\exp(\beta_{1A}) = 1.5$, $\tau_F = 0.5$, $\beta_1^\circ = 0$, and non-inferiority margin $\exp(\beta_{1M}) = 2$; verification by calculation of the empirical type I error and the empirical power; $n_{sim} = 1000$.

Parameter setting					Analytical results				Empirical results (fixed n) [†]			Empirical results (fixed τ_A) [†]			
N_{pop}	Target power (%)	λ	ϕ	ρ	κ	ν	τ_A	$\mathbb{E}\{N(\tau_A)\}$	$\hat{\tau}_A$	Emp. power (%)	Emp. type I error (%)	$\hat{\mathbb{E}}\{N(\tau_A)\}$	Emp. power (%)	Emp. type I error (%)	
40,000	80	10	0.80	0.005	8.58	34.13	0.059	15306	0.059	78.0	4.4	15305	80.1	4.3	
				0.010	8.49	33.53	0.065	16311	0.065	78.4	4.3	16311	79.7	6.5	
				1.60	0.005	11.02	76.30	0.050	12223	0.050	78.7	4.2	12228	78.7	5.6
		20	0.80	0.010	10.74	73.42	0.058	13480	0.058	81.8	4.4	13478	80.0	5.2	
				0.005	17.55	139.34	0.012	7934	0.012	77.8	4.6	7935	77.8	6.1	
				0.010	17.38	136.86	0.014	8880	0.014	78.4	5.3	8879	78.8	6.0	
	90	10	0.80	1.60	0.005	23.89	350.39	0.010	6173	0.010	78.1	6.1	6173	78.8	5.1
				0.010	23.35	336.74	0.012	7283	0.012	79.9	3.9	7289	77.3	5.1	
				0.010	7.81	29.87	0.125	23194	0.125	89.4	4.4	23192	89.8	4.5	
		20	0.80	0.010	7.65	29.10	0.147	24854	0.147	89.6	4.6	24853	87.5	4.8	
				1.60	0.005	9.43	62.12	0.121	19616	0.121	88.4	5.9	19617	88.4	5.4
				0.010	8.95	58.76	0.162	22024	0.162	90.4	4.6	22026	88.7	5.4	
	60,000	80	10	0.80	0.005	9.12	37.79	0.031	14449	0.031	78.2	5.1	14448	76.4	4.1
					0.010	9.06	37.36	0.033	15390	0.033	76.9	4.9	15390	79.0	3.4
					1.60	0.005	12.14	89.24	0.024	11182	0.024	80.9	6.6	11179	78.2
			20	0.80	0.010	11.96	87.04	0.028	12277	0.028	79.7	3.7	12279	81.0	5.6
					0.005	18.06	146.89	0.007	7741	0.007	79.1	6.5	7739	79.8	5.1
					0.010	17.95	145.19	0.008	8657	0.008	81.2	4.7	8660	80.3	6.3
90		10	0.80	1.60	0.005	24.98	379.99	0.006	5948	0.006	78.3	5.6	5947	80.9	4.9
				0.010	24.63	370.30	0.007	6991	0.007	78.5	3.7	6992	78.5	5.4	
				0.010	8.71	34.97	0.051	20904	0.051	89.2	5.0	20907	89.7	5.1	
		20	0.80	0.010	8.63	34.42	0.056	22259	0.056	89.3	6.7	22252	90.2	6.2	
				1.60	0.005	11.29	79.22	0.042	16550	0.042	90.4	4.9	16551	90.9	4.3
				0.010	11.04	76.46	0.049	18231	0.049	89.8	4.4	18230	89.5	5.5	

[†] Columns headed “fixed n ” recruited donors until the expected number of donors $\mathbb{E}\{N(\tau_A)\}$ were recruited and at that point the recruitment period ended (and so τ_A is random). The average duration of the random recruitment period $\hat{\tau}_A$ was very close to τ_A . For columns headed “fixed τ_A ” the accrual period was fixed at the derived τ_A and the actual number of donors recruited varies across the $n_{sim} = 1000$ simulations for each row of the table.

be generated using the pre-specified marginal means μ_0 and μ_1 , and pairwise association ρ . The treatment effect $\hat{\beta}_1$ is estimated by solving the generalized estimating equation in (3.2) and the hypothesis test is based on $\hat{\beta}_1$ and its robust variance estimated as described in Section 3.2. We repeated this procedure for $n_{sim} = 1000$ simulated datasets for both superiority (Table 3.1) and non-inferiority (Table 3.2) designs. For the “fixed n ” design we report the empirical mean $\hat{\tau}_A$ and for the “fixed τ_A ” design we report the mean sample size as $\hat{\mathbb{E}}(N(\tau_A))$.

The simulations reveal that the mean duration of the empirical recruitment periods under in very close agreement with the derived one τ_A when we treat the recruitment duration as random (columns 8 and 10 of Table 3.1 and 3.2), and the average number of the empirical sample sizes is close to the expected sample size $\mathbb{E}\{N(\tau_A)\}$ when it is random (columns 9 and 13 of Table 3.1 and 3.2) under the fixed τ_A implementation. The empirical power for each scenario are summarized in Table 3.1 and 3.2 and are found to be in close agreement with the nominal power set at the design stage for both the 80% and 90% power settings. Finally we note that there is good control of the type I error rate (rejection rate under H_0) at the 5% level for both settings affirming that the robust variance estimates adequately address the misspecification of the variance function implicit in the use of the log link and the clustering in the observations within individuals over successive donations.

3.4.3 Design of a Plasma Donation Study

The IMproving PlasmA CollecTion trial (IMPACT) was a randomized controlled trial designed to assess the non-inferiority of a novel “percent plasma nomogram” (PPN) for plasma collection compared to the standard nomogram in terms of the incidence of significant hypotensive adverse events (Hartmann et al, 2021). The PPN nomogram was developed to personalize the target volume of plasma to be withdrawn from a donor; for some donors less plasma may be withdrawn while for others more plasma might be targeted compared to the volume under the established nomogram. It recruited donors from three plasma collection centers (i.e. Charlotte, NC; Spokane, WA; Wichita, KS) between January 6 and March 26, 2020. The primary endpoint of the trial is the occurrence of a significant hypotensive adverse event, defined according to the modified IQPP classification (see supporting information in Hartmann et al (2021)). Here we illustrate how the proposed design framework can be employed in this context for planning a plasma donation study.

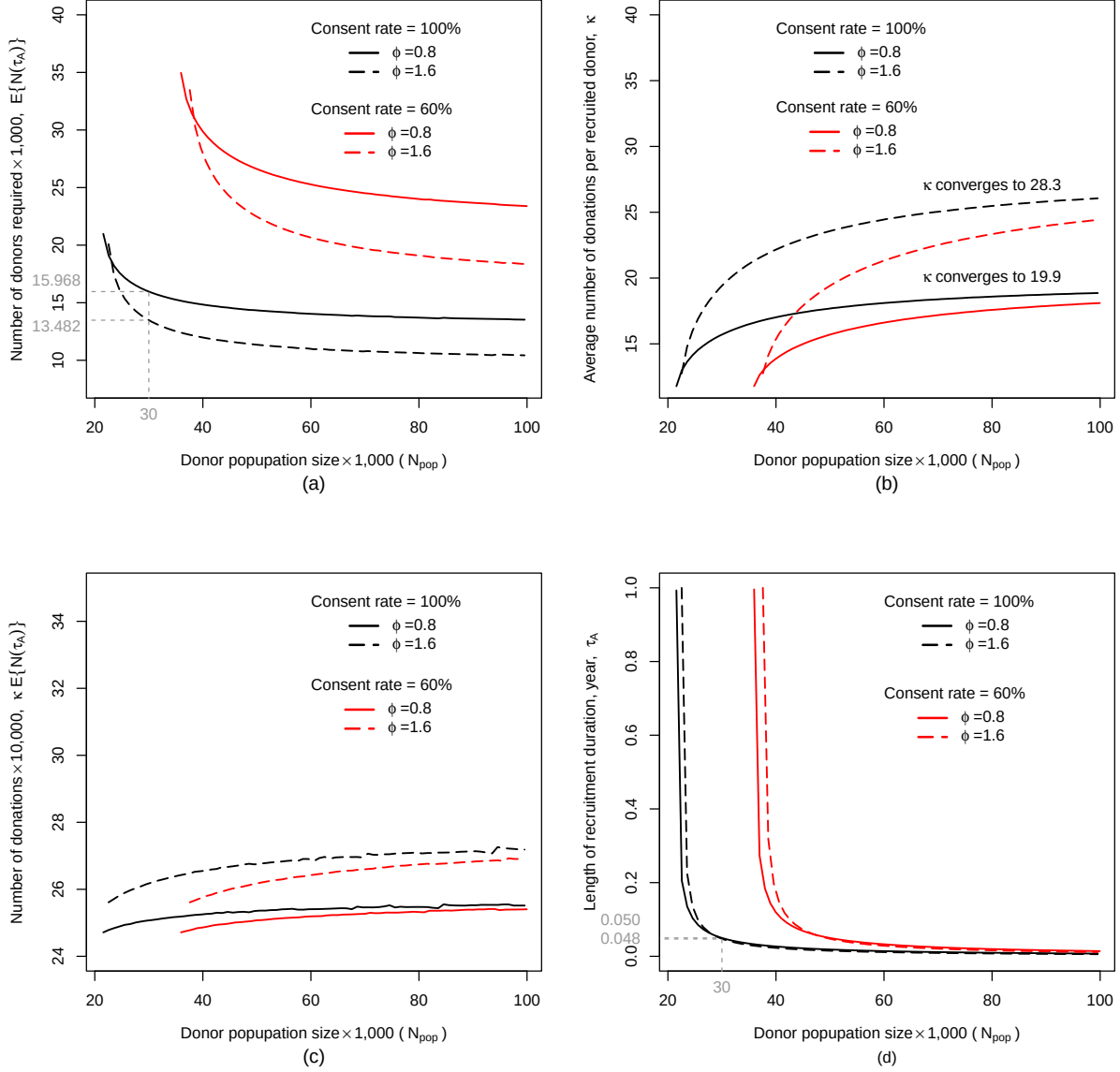


Figure 3.2: Impact of donor population size (N_{pop}) on the number of donors required for recruitment ($E\{N(\tau_A)\}$), average number of donations per donor (κ), total number of donations to be collected ($\kappa E\{N(\tau_A)\}$), and recruitment duration (τ_A) in a non-inferiority trial with $\lambda = 21$, $\mu_0 = 0.035\%$, $\exp(\beta_{1A}) = 1.5$, $\exp(\beta_{1M}) = 2.5$, $\rho = 0.005$, $\phi = 0.8, 1.6$, $\tau_F = 0.5$, $\alpha = 5\%$, target power=80%, and consent rate=60%, 100%.

Modelling our design based on the goals of the IMPACT trial we suppose individual donors are randomized to provide donations using the PPN or standard nomograms during the follow-up phase with 50:50 allocation to the two arms (i.e. we set $P(X_i = 1) = 0.5$). Based on Canadian data on established plasma donors the expected annual number of donations is approximately 21 so we set $\lambda = 21$. Using the standard nomogram the significant hypotensive adverse event rate per donation under the control arm is set to $\mu_0 = 0.035\%$. We suppose the experimental donation algorithm aims to draw more plasma from some donors and hence accommodate a 50% increase in the risk of events in the experimental arm and set $\exp(\beta_{1A}) = 1.50$. Given the event rate is very low, we set a non-inferiority margin of $\exp(\beta_{1M}) = 2.5$, and consider trial planning with $H_0 : \beta_1 \geq \beta_{1M}$ versus $H_A : \beta_1 < \beta_{1M}$ with significant level $\alpha = 5\%$ and target power at 80%. We further assume an exchangeable within-donor serial correlation in the responses over successive donations of $\rho = 0.005$. We investigate settings with $\phi = 0.80$ and 1.6 for moderate and great heterogeneity in the donation rate in the donor population.

Figure 3.2(a) shows the relationship between the size of the donor population (horizontal axis) and the expected total number of donors recruited to the study (vertical axis) for settings with a 100% consent rate and a 60% consent rate along with the two values of ϕ . We give this figure to point out the implications of engaging more donation centers in a study, which is an effective way of increasing the size of the study donor population and shortening the duration of the accrual period. With a 6-month follow-up period ($\tau_F = 0.5$) and total donor population of size $N_{\text{pop}} = 30,000$ across all sites, for example, we find that the accrual period should be about 0.050 years (or 2.6 weeks) or 0.048 years (about 2.5 weeks) for $\phi = 0.8$ and 1.6 settings respectively. With a 100% consent rate, these values of τ_A lead to an expected total of 15,968 donors to be recruited by the end of the accrual period to achieve 80% power in a non-inferiority study with $\phi = 0.8$, and a slight decrease in the expected number of donors to 13,482 if $\phi = 1.6$; see the lower left corner of Figure 3.2(a) where these are marked by grey lines. The greater variance ϕ results in fewer expected donors because the sample in such settings is enriched with high frequency donors - that is, the recruited donors will have a higher donation rate on average. More details on the relation between the size of the donor population and the duration of the accrual period are available from Figure 3.2(d). Figure 3.2(b) illustrates the selection effect more clearly, where it can be seen that the average number of donations per recruited donor increases with the donor population size. As we open more sites and the donor population size increases, the average number of donations per donor (κ) converges to $\lambda\tau_F(\phi + 1) + 1$ which can be obtained from (3.12). A lower consent rate has an interesting effect in this setting in that it leads to an extension of the recruitment period, and the recruitment of more lower-frequency donors, thereby reducing the average number of donations per

donor as shown in Figure 3.2(b). Figure 3.2(c) shows how the total expected number of donations on study varies as a function of population size for different ϕ values and consent rates. The expected total number of donations shown in Figure 3.2(c) is useful for resource planning, e.g., determining how many plasma kits should be available for the study. While it generally increases with the donor population, it does not increase monotonically. This is because the total number of donations is the product of the growing sample size and the decreasing average number of donations per recruited donor.

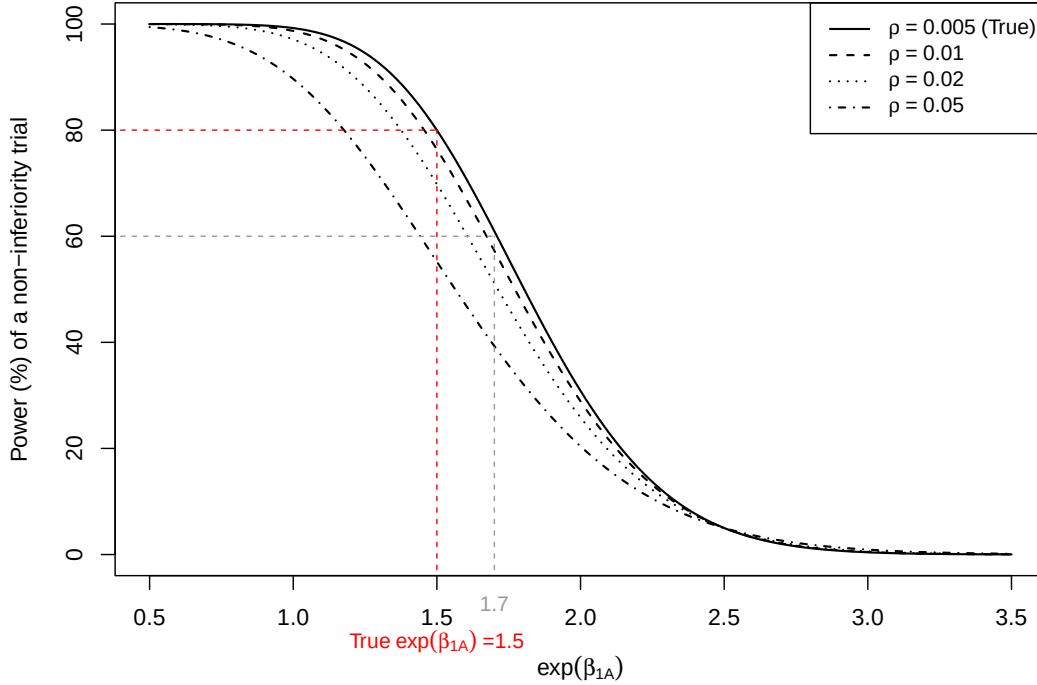


Figure 3.3: Power of a non-inferiority trial when $\exp(\beta_{1A})$ and ρ are misspecified; true parameters are set to $N = 40,000$, $\lambda = 21$, $\mu_0 = 0.035\%$, $\exp(\beta_{1A}) = 1.5$, $\exp(\beta_{1M}) = 2.5$, $\rho = 0.005$, $\phi = 0.8$, $\tau_F = 0.5$, $\alpha = 5\%$, and target power=80%.

The sample size requirements are derived from the large number of models and parameters so it is important to assess the sensitivity of the design to misspecification of the design parameters. Here we assess the impact of the misspecification of the correlation of responses (ρ) and the treatment effect (β_{1A}). If the true treatment effect β_{1A} is lower or

greater than was specified at the planning stage then there will be a great or lower power for demonstrating non-inferiority respectively; this is illustrated in the power curves of Figure 3.3. If, for example, $\exp(\beta_{1A})$ is equal to 1.7 rather than the value of 1.5 influencing the design, then there is an appreciable drop in power from 80% to 60% (assuming ρ was correctly specified at $\rho = 0.005$). Additional power curves are provided in Figure 3.3 to show the impact of under-estimating the within-donor correlation in the binary responses. Here it can be seen that larger values of ρ will decrease the power of the study. This suggests the importance of historical data to ensure suitable values of ρ are used at the planning stage. Adaptive designs can alternatively be used to adjust the duration of the accrual period or sample size accordingly.

3.5 Remarks

We have developed a statistical framework for the plasma donation trial planning and emphasized sample size calculations in terms of the random number of recruited individuals, with the duration of the follow-up phase and population size set a priori to determine how long the recruitment phase should be (Cook, 1995). Interestingly as the duration of the recruitment phase increases the effect of zero-truncation on the count during this phase is reduced and the distribution of U in the target population more closely aligns with the population distribution. Shortening the duration of the recruitment phase will mean that selected individuals donate at a higher rate than the typical donor in the population and so furnish more data on the response to donations and the intervention effect as shown in Figure 3.2.

We have considered superiority trials and non-inferiority trials and shown that the large sample results offer useful guidance in both settings. Sequential monitoring of trials of this sort raises interesting challenges. Specifically the individuals in the population who are recruited earlier during the recruitment phase will tend to be more frequent donors. As a result they will have a higher random effect and donate more during the treatment phase. Individuals who are recruited later during the recruitment phase as those that have not been encountered previously and so are more infrequent donors. As a result they will tend to experience fewer donations during the treatment phase. This phenomenon means that information on recruitment time is acquired in studies with this type of design much more rapidly in the early part of the trial compared to later. This has an impact on planning the timing of interim analyses which we explore in the following sections.

3.6 Sequential Analysis of Marked Point Processes

Large clinical trials are now routinely designed to incorporate interim analyses to monitor the safety and effectiveness of experimental interventions. Early termination of trials showing unexpected adverse effects of a new treatment minimizes exposure to harm, while early termination when treatments are found to be effective facilitates rapid dissemination of findings and deployment of more useful treatments to affected individuals. In addition to these ethical reasons, trials with planned interim analyses can be expected to reduce costs due to early termination when large treatment effects are seen since fewer participants may need to be recruited and follow-up can be shortened of enrolled participants. The statistical challenge involves control of the experimental type I error rate when conducting repeated significance tests on data accumulating over the course of the study ([Armitage et al, 1969](#)).

In early work on group sequential trials, [Pocock \(1977\)](#) proposed controlling the type I error rate through derivation of a single critical value, which can be used to guide stopping decisions and determine the statistical significance of the trial results. [O'Brien and Fleming \(1979\)](#) later proposed a design wherein different critical values could be used at successive interim analyses with greater thresholds for early analyses leading to designs which are more conservative for early stopping. These methods all require specification of a fixed number of interim analyses timed at roughly equal amounts of information under the assumption that data accumulated between successive analysis times are independent of the past data. The so-called independent increment structure enables the use of a recursive integration method ([Armitage et al, 1969](#); [Reboussin et al, 2000](#)) to compute standard group sequential boundaries for repeated significance tests. [Slud and Wei \(1982\)](#) generalized the approach to sequential monitoring to allow for departures from the independent increments assumption, greatly expanding the range of analyses for which group sequential analyses could be planned. [Lan and DeMets \(1983\)](#) proposed a more general concept of an alpha spending function which enables a much more flexible and adaptive way to compute critical values for interim tests where the number and timing of analyses can be modified during the course of the study. In trials involving responses based on repeated measures, the independent increments structure for successive test statistics is violated because of the serial dependence of the responses within individuals. [Lee and Demets \(1991\)](#) proposed to address this for continuous responses based on random-effects models while [Wei et al \(1990\)](#) and [Lee et al \(1996\)](#) used generalized estimating equations ([Liang and Zeger, 1986](#)) to develop a framework applicable to a broader range of data types. Many of these methods however are based on designs in which the repeated measurements are taken at specified fixed time points following randomization.

In many areas interest lies in assessing the effect of an experimental device or intervention which is delivered repeatedly over time but at times determined by the individual and hence which cannot be specified a priori. Examples are ubiquitous and arise in neurology when interest lies in the evaluation of therapies for acute migraine which are only administered when migraines develop ([Pascual et al, 2000](#)), in respirology where bronchodilators are administered for the resolution of symptoms in recurrent exacerbations in asthma ([Wilson et al, 2004](#)), and in transfusion medicine where therapeutic transfusion of platelet or red cell products are only given when the need is identified (i.e. platelets are transfused when thrombocytopenia is sufficiently severe that patients are at risk of serious bleeds ([Webert et al, 2006](#))). A similar data structure arises in blood donation research in which healthy donors give blood repeatedly over time. Here interest may lie in studying the occurrence of adverse events upon blood donation - in particular in comparing whether two methods of blood collection lead to similar adverse event profiles in donors. In all such settings one can think of the data as arising from a marked point process where the point is the time an intervention is administered and the mark is the response to the intervention. We consider the setting where individuals are randomized to one of two intervention arms and will receive the assigned intervention whenever it is required during the course of follow-up. The work in this Chapter is motivated by the setting of a blood donation study so here the points are the times that donors arrive at a donation center to give blood and the mark is an indicator of whether an adverse reaction occurred.

We define notation and give the generalized estimating functions used for estimation and large sample theory for analyses in Section 3.7. The joint distribution of the sequential test statistics is then derived. In Section 3.8 an illustrative simple setting is studied in detail where there is no covariate adjustment and the marginal models for treatment effect simply involve the binary treatment indicator. The precise forms of the large sample distributions are given which identify the terms needed to facilitate sample size calculations for a two-arm clinical trial. One element of this large sample distribution involves the distribution of the number of transfusions during the follow-up period. Section 3.9 proposes the use of a zero-truncated negative binomial process for the need for transfusions. We consider a nuance whereby donors giving more frequent blood donations will tend to be encountered sooner than those donating less frequently. In probabilistic terms, we allow the intensity for transfusions to be higher for those individuals who are recruited earlier in the course of the accrual period - the impact of this is explored in Section 3.9. Alpha spending functions are reviewed in Section 3.10. Sample size calculations for the study with interim analyses are given. Both superiority and non-inferiority study design are considered in Section 3.10.2 and Section 3.10.3 respectively. Section 3.11 gives a simulation study demonstrating the good control of the type I error rate and the power for the derived sample sizes. In Section

3.11.2 we illustrate an example of a plasma donation study design with interim analyses. Discussion and concluding remarks are given in Section 3.12.

3.7 Statistical Formulation and Methods

3.7.1 Notation and Generalized Estimating Equations

Let $\mathcal{K}_i(t)$ denote the cumulative number of donations over a window of calendar time $(0, t]$ for individual i in the donor population of interest. Let $\{\mathcal{K}_i(t), 0 < t\}$ be a right-continuous counting process that records the cumulative number of donations from individual i over the period of calendar time $(0, t]$ with $d\mathcal{K}_i(t) = \mathcal{K}_i(t) - \mathcal{K}_i(t^-) = 1$ indicating a donation took place at calendar time t , and $d\mathcal{K}_i(t) = 0$ otherwise. With the history of donations over $(0, t)$ given by $\mathcal{H}_i(t) = \{d\mathcal{K}_i(v), 0 < v < t\}$, we assume donations occur in the population according to a mixed Poisson process with conditional intensity (Cook and Lawless, 2007)

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta\mathcal{K}_i(t) = 1 \mid \mathcal{H}_i(t), U_i = u_i)}{\Delta t} = u_i \lambda(t) , \quad (3.17)$$

where $\Delta\mathcal{K}_i(t) = \mathcal{K}_i(t + \Delta t^-) - \mathcal{K}_i(t)$ is the number of donations over the interval $[t, t + \Delta t)$, $U_i > 0$ is a random effect with mean one and variance $\phi > 0$, and $\lambda(t)$ is a baseline population average rate function.

We consider a study with an accrual period of length τ_A and a follow-up period for each recruited individual of length τ_F making the total study duration $\tau = \tau_A + \tau_F$. It is helpful to consider two time-scales with t denoting the time from the start of the accrual period of the trial with $0 \leq t \leq \tau$ and s the time from recruitment for an individual with $0 \leq s \leq \tau_F$. We may refer to the former as “trial time” (which is the calendar time with the time origin for t defined as the start of the accrual period), and the latter as “study time” (for time on-study for an individual). If L_i is the random recruitment time for individual i and l_i is its realization then provided $0 \leq l_i \leq \tau_A$ individual i has been on study for a length of time s at $t = l_i + s$. Figure 3.4 contains a Lexis diagram depicting the accrual and study follow-up periods which define the trial duration, as well as the relation between time-scales t on the horizontal axis and s on the vertical axis. The 45° lines emanate from the horizontal axis at the time of recruitment with 10 hypothetical individuals depicted. The donation process in (3.17) plays an important role in study planning as it determines the distribution of the random accrual time L_i as well as the associated event intensity during the study period among recruited individuals. In brief the heterogeneity in the rate

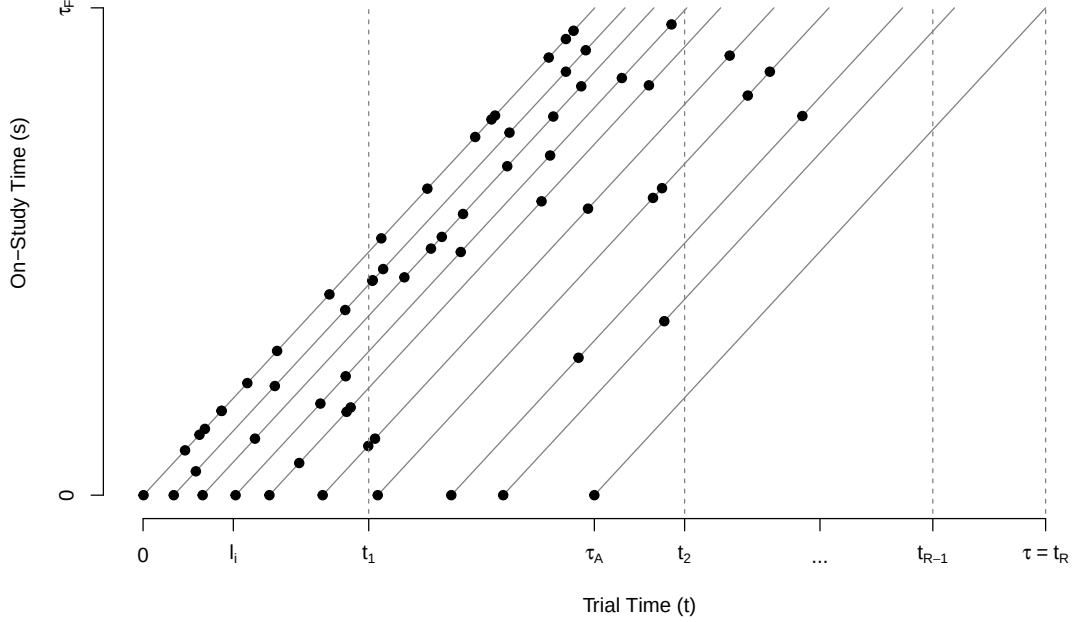


Figure 3.4: An illustration of the recruitment process for 10 subjects along with their repeated measurements. Interim analyses are planned at study times $t_1, t_2, \dots, t_{R-1}$, and a final analysis at t_R . The length of the accrual period is τ_A , the length of the follow-up period is τ_F for each individual, and the entire study duration is $\tau = \tau_A + \tau_F$.

of donations among individuals in the donor population means that frequent donors will be encountered and recruited early during the accrual period, and as the time from the start of the accrual period passes more donors with lower donation rates will be encountered. This is depicted in Figure 3.4 by the more frequent donations (represented by the dots on the 45° lines) in the individuals recruited earlier compared to individuals recruited later in the period $(0, \tau_A]$. We consider the implications of this in more detail in Section 3.9.

Consider the setting where n consenting donors are recruited and randomized upon encountering them during the accrual period. The total number of donations received from the recruited individual i by time t from the start of the trial is

$$K_i(t) = \int_0^t Y_i(v) d\mathcal{K}_i(v)$$

where $Y_i(t) = I(l_i \leq t \leq l_i + \tau_F)$. Let Y_{ik} be the response measured at the k th donation during study follow-up which is observed at time t if $K_i(t^-) = k - 1$ and $dK_i(t) = 1$; we can refer to this as the mark associated with the k th donation and here we consider the case where Y_{ik} is a binary response indicating the occurrence of an adverse event where $Y_{ik} = 1$ denotes that an adverse event occurs and $Y_{ik} = 0$ otherwise. Ultimately we may observe the set of marks $\{Y_{ik}, k = 1, \dots, K_i(\tau)\}$ for individual i , $i = 1, \dots, n$. Let X_i indicate the treatment arm where $X_i = 1$ if individual i is assigned to the experimental arm and $X_i = 0$ otherwise. If $\mathbf{V}_{ik}$ represents a $(p - 2) \times 1$ covariate vector measured at the time of the k th donation we let $\mathbf{Z}_{ik} = (1, X_i, \mathbf{V}_{ik}')'$ be the $p \times 1$ covariate vector, $k = 1, 2, \dots, K_i(\tau)$, $i = 1, \dots, n$.

Suppose interim analyses are planned at trial times t_r , $r = 1, \dots, R - 1$ with a final analysis at study completion (i.e. $t_R = \tau$); see Figure 3.4. Let $\Delta_i(t_r) = I(l_i \leq t_r)$ indicate that individual i entered the study by time t_r and $n_r = \sum_{i=1}^n \Delta_i(t_r)$ denote the total number of such individuals contributing to the r th analysis, $r = 1, \dots, R$, with $n = n_R$. Specifically the vector of responses $\mathbf{Y}_i(t_r) = (Y_{i1}, \dots, Y_{iK_i(t_r)})'$ is the set of indicators from the $K_i(t_r)$ donations observed by the time of the r th analysis with $\mathbf{Z}_i(t_r) = (\mathbf{Z}_{i1}', \dots, \mathbf{Z}_{iK_i(t_r)}')'$ the corresponding matrix of covariates. The generalized estimating equations (Liang and Zeger, 1986) based on subjects who have been recruited by the time of the r th interim analysis can be written as

$$\begin{aligned} U(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha}) &= \sum_{i=1}^n U_i(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha}) \\ &= \sum_{i=1}^n \Delta_i(t_r) D_i'(t_r; \boldsymbol{\beta}) \Sigma_i(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} \{\mathbf{Y}_i(t_r) - \boldsymbol{\mu}_i(t_r; \boldsymbol{\beta})\} = 0, \end{aligned} \quad (3.18)$$

where $\boldsymbol{\mu}_i(t_r; \boldsymbol{\beta}) = (\mu(\mathbf{Z}_{i1}; \boldsymbol{\beta}), \dots, \mu(\mathbf{Z}_{iK_i(t_r)}; \boldsymbol{\beta}))'$ is the $K_i(t_r) \times 1$ vector of means with

$$\mu(\mathbf{Z}_{ik}; \boldsymbol{\beta}) = \mathbb{E}(Y_{ik} \mid \mathbf{Z}_{ik}; \boldsymbol{\beta}) = P(Y_{ik} = 1 \mid \mathbf{Z}_{ik}; \boldsymbol{\beta}), \quad k = 1, 2, \dots, K_i(t_r),$$

indexed by $\boldsymbol{\beta}$, $D_i(t_r; \boldsymbol{\beta}) = \partial \boldsymbol{\mu}_i(t_r; \boldsymbol{\beta}) / \partial \boldsymbol{\beta}'$, and $\Sigma_i(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha}) = \text{cov}(\mathbf{Y}_i(t_r) \mid \mathbf{Z}_i(t_r))$ is the working covariance matrix indexed by a parameter vector $\boldsymbol{\alpha}$. The marginal probability that $Y_{ik} = 1$ is very small in the motivating study since serious hypotensive adverse events are very rare in blood donation as we described in Section 3.2.1 (Li and Cook, 2024b). The natural log link function can therefore be specified under the ‘‘Poisson’’ approximation giving the generalized linear model (McCullagh, 2019) of the form

$$\log(\mu(\mathbf{Z}_{ik}; \boldsymbol{\beta})) = \beta_0 + \beta_1 X_i + \boldsymbol{\beta}_2' \mathbf{V}_{ik} = \boldsymbol{\beta}' \mathbf{Z}_{ik},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \boldsymbol{\beta}'_2)'$. This gives

$$\beta_1 = \log \frac{P(Y_{ik} = 1 \mid X_i = 1, \mathbf{V}_{ik}; \boldsymbol{\beta})}{P(Y_{ik} = 1 \mid X_i = 0, \mathbf{V}_{ik}; \boldsymbol{\beta})}$$

as the log relative rate of an adverse event for the experimental arm compared to the control arm conditional on the covariate $\mathbf{V}_{ik}$ but we note that if $\mathbf{V}_{ik} \perp X_i$ then this retains the same interpretation if we do not adjust for auxiliary covariates by setting $\boldsymbol{\beta}_2 = \mathbf{0}$.

The covariance parameter $\boldsymbol{\alpha}$ can be estimated using the moment estimator based on Pearson residuals (Liang and Zeger, 1986), or through second-order estimating equations (Prentice, 1988; Zhao and Prentice, 1990; Prentice and Zhao, 1991). Liang and Zeger (1986) showed that $\mathbf{U}(t_r; \boldsymbol{\beta}, \hat{\boldsymbol{\alpha}})$ is asymptotically equivalent to $\mathbf{U}(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha})$ if $\hat{\boldsymbol{\alpha}}$ is $\sqrt{n}$ -consistent so $\boldsymbol{\alpha}$ will be estimated based on the data accumulated by study time t_r and updated at each interim analysis. By the Central Limit Theorem,

$$\frac{1}{\sqrt{n}} \mathbf{U}(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha}) \xrightarrow{d} \text{MVN}(0, \mathcal{B}(t_r)) \quad , \quad \text{as } n \rightarrow \infty \quad (3.19)$$

where

$$\mathcal{B}(t_r) = \text{cov}(\mathbf{U}_i(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha}), \mathbf{U}_i(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha})) = \mathbb{E} \{ \mathbf{U}_i(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha}) \mathbf{U}_i'(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha}) \} \quad (3.20)$$

is the $p \times p$ asymptotic covariance matrix.

The solution to (3.18) is the estimate of $\boldsymbol{\beta}$ at the r th interim analysis and is denoted by $\hat{\boldsymbol{\beta}}(t_r)$. Our interest lies primarily in $\hat{\beta}_1(t_r)$, the estimated coefficient of X_i , since $\exp(\hat{\beta}_1(t_r))$ is an estimate of the relative rate of an adverse event in the experimental to control arms. Note

$$\sqrt{n} \left(\hat{\boldsymbol{\beta}}(t_r) - \boldsymbol{\beta} \right) \xrightarrow{d} \text{MVN}(0, \Omega(t_r)) \quad \text{as } n \rightarrow \infty \quad , \quad (3.21)$$

where the robust covariance matrix is

$$\Omega(t_r) = \mathcal{A}^{-1}(t_r) \mathcal{B}(t_r) \{ \mathcal{A}^{-1}(t_r) \}' \quad ,$$

and

$$\mathcal{A}(t_r) = -\mathbb{E} \left\{ \frac{\partial}{\partial \boldsymbol{\beta}'} \mathbf{U}_i(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha}) \right\} = \mathbb{E} \left\{ \Delta_i(t_r) D_i'(t_r; \boldsymbol{\beta}) \Sigma_i(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} D_i(t_r; \boldsymbol{\beta}) \right\} \quad . \quad (3.22)$$

3.7.2 Joint Distribution of Sequential Test Statistics

Here we are interested in the effect of the experimental device and aim to test

$$H_0 : \beta_1 = \beta_1^\circ \text{ versus } H_A : \beta_1 \neq \beta_1^\circ .$$

We consider the group sequential setting where analyses are planned at $t_1 < t_2 < \dots < t_R = \tau$. The p -dimensional generalized estimating functions $\mathbf{U}(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha})$ in (3.18) can lead to a pR -dimensional joint multivariate normal distribution of $\{n^{-\frac{1}{2}}\mathbf{U}(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha}), r = 1, \dots, R\}$ (Lee et al, 1996). Since our primary interest is β_1 and for ease of derivation of the joint distribution of sequential test statistics, we adopt the approach proposed by Lee et al (1996) in which $\mathbf{U}(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha})$ is partitioned as

$$(U_{\beta_1}(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha}), \mathbf{U}'_{\boldsymbol{\theta}}(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha}))' ,$$

where $U_{\beta_1}(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha})$ is the one-dimensional estimating function for β_1 , and $\boldsymbol{\theta} = (\beta_0, \beta_2)'$ contains the remaining regression parameters (i.e. excluding β_1). Then $\mathbf{U}_{\boldsymbol{\theta}}(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha})$ is the $(p-1)$ -dimensional estimating function with respect to $\boldsymbol{\theta}$. Lee et al (1996) proposed using “profile score function” or profile estimating function of β_1 to construct the sequential test statistics and derive its asymptotic joint distribution. The profile estimating function (Lee et al, 1996; Rotnitzky and Jewell, 1990) is written as

$$\bar{U}(t_r; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha}) = U_{\beta_1}(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha}) - \mathcal{A}_{\beta_1\boldsymbol{\theta}}(t_r)\mathcal{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(t_r)\mathbf{U}_{\boldsymbol{\theta}}(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha}) ,$$

where $\mathcal{A}_{\beta_1\beta_1}(t_r)$, $\mathcal{A}_{\beta_1\boldsymbol{\theta}}(t_r)$, $\mathcal{A}_{\boldsymbol{\theta}\beta_1}(t_r)$, and $\mathcal{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}(t_r)$ are submatrices defined conformably with $(\beta_1, \boldsymbol{\theta})'$ as

$$\mathcal{A}(t_r) = \begin{pmatrix} \mathcal{A}_{\beta_1\beta_1}(t_r) & \mathcal{A}_{\beta_1\boldsymbol{\theta}}(t_r) \\ \mathcal{A}_{\boldsymbol{\theta}\beta_1}(t_r) & \mathcal{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}(t_r) \end{pmatrix} .$$

By asymptotic multivariate normality and using a linear transformation of (3.19) we have

$$\frac{1}{\sqrt{n}} \bar{U}(t_r; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha}) \xrightarrow{d} N(0, M_1(t_r)) , \quad \text{as } n \rightarrow \infty \quad (3.23)$$

with asymptotic variance

$$M_1(t_r) = \mathcal{B}_{\beta_1\beta_1}(t_r) + \mathcal{A}_{\beta_1\boldsymbol{\theta}}(t_r)\mathcal{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(t_r)\mathcal{B}_{\boldsymbol{\theta}\boldsymbol{\theta}}(t_r)[\mathcal{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(t_r)]'[\mathcal{A}_{\beta_1\boldsymbol{\theta}}(t_r)]' - 2\mathcal{A}_{\beta_1\boldsymbol{\theta}}(t_r)\mathcal{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(t_r)\mathcal{B}_{\boldsymbol{\theta}\beta_1}(t_r) ,$$

where $\mathcal{B}_{\beta_1\beta_1}(t_r)$, $\mathcal{B}_{\beta_1\boldsymbol{\theta}}(t_r)$, $\mathcal{B}_{\boldsymbol{\theta}\beta_1}(t_r)$, and $\mathcal{B}_{\boldsymbol{\theta}\boldsymbol{\theta}}(t_r)$ are submatrices of $\mathcal{B}(t_r)$, and

$$\mathcal{B}(t_r) = \begin{pmatrix} \mathcal{B}_{\beta_1\beta_1}(t_r) & \mathcal{B}_{\beta_1\boldsymbol{\theta}}(t_r) \\ \mathcal{B}_{\boldsymbol{\theta}\beta_1}(t_r) & \mathcal{B}_{\boldsymbol{\theta}\boldsymbol{\theta}}(t_r) \end{pmatrix}.$$

The asymptotic variance $M_1(t_r)$ can be simplified to $\mathcal{A}_{\beta_1\beta_1}(t_r) - \mathcal{A}_{\beta_1\boldsymbol{\theta}}(t_r)\mathcal{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(t_r)\mathcal{A}_{\boldsymbol{\theta}\beta_1}(t_r)$ if the working covariance matrix $\Sigma_i(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha})$ is correctly specified since in this case $\mathcal{A}(t_r) = \mathcal{B}(t_r)$.

The R -dimensional vector of profile estimating functions over all analyses times,

$$\left\{ \frac{1}{\sqrt{n}} \bar{U}(t_r; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha}), r = 1, \dots, R \right\}, \quad (3.24)$$

is asymptotically multivariate normal with zero mean and covariance at any two study times t_r and t_s is given by

$$\begin{aligned} M_2(t_r, t_s) &= \text{cov}(\bar{U}_i(t_r; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha}), \bar{U}_i(t_s; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha})) \\ &= \mathcal{B}_{\beta_1\beta_1}(t_r, t_s) + \mathcal{A}_{\beta_1\boldsymbol{\theta}}(t_r)\mathcal{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(t_r)\mathcal{B}_{\boldsymbol{\theta}\boldsymbol{\theta}}(t_r, t_s)[\mathcal{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(t_s)]'[\mathcal{A}_{\beta_1\boldsymbol{\theta}}(t_s)]' \\ &\quad - \mathcal{B}_{\beta_1\boldsymbol{\theta}}(t_r, t_s)[\mathcal{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(t_s)]'[\mathcal{A}_{\beta_1\boldsymbol{\theta}}(t_s)]' - \mathcal{A}_{\beta_1\boldsymbol{\theta}}(t_r)\mathcal{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(t_r)\mathcal{B}_{\boldsymbol{\theta}\beta_1}(t_r, t_s), \end{aligned} \quad (3.25)$$

where submatrices $\mathcal{B}_{\beta_1\beta_1}(t_r, t_s)$, $\mathcal{B}_{\beta_1\boldsymbol{\theta}}(t_r, t_s)$, $\mathcal{B}_{\boldsymbol{\theta}\beta_1}(t_r, t_s)$, and $\mathcal{B}_{\boldsymbol{\theta}\boldsymbol{\theta}}(t_r, t_s)$ are from

$$\mathcal{B}(t_r, t_s) = \mathbb{E} \{ \mathbf{U}_i(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha}) \mathbf{U}_i'(t_s; \boldsymbol{\beta}, \boldsymbol{\alpha}) \} = \begin{pmatrix} \mathcal{B}_{\beta_1\beta_1}(t_r, t_s) & \mathcal{B}_{\beta_1\boldsymbol{\theta}}(t_r, t_s) \\ \mathcal{B}_{\boldsymbol{\theta}\beta_1}(t_r, t_s) & \mathcal{B}_{\boldsymbol{\theta}\boldsymbol{\theta}}(t_r, t_s) \end{pmatrix}. \quad (3.26)$$

By Taylor expansion $\sqrt{n}(\widehat{\beta}_1(t_r) - \beta_1^\circ)$ can be written as

$$\sqrt{n} \left(\widehat{\beta}_1(t_r) - \beta_1^\circ \right) = \left\{ -\frac{1}{n} \frac{\partial \bar{U}(t_r; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha})}{\partial \beta_1} \right\}^{-1} \frac{1}{\sqrt{n}} \bar{U}(t_r; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha}). \quad (3.27)$$

Under the null hypothesis $\beta_1 = \beta_1^\circ$, (3.23) and (3.27) lead to

$$\sqrt{n} \left(\widehat{\beta}_1(t_r) - \beta_1^\circ \right) \xrightarrow{d} N \left(0, \frac{M_1(t_r)}{M_3^2(t_r)} \right), \quad \text{as } n \rightarrow \infty \quad (3.28)$$

where

$$M_3(t_r) = \mathbb{E} \left\{ -\frac{\partial \bar{U}_i(t_r; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha})}{\partial \beta_1} \right\} = \mathcal{A}_{\beta_1\beta_1}(t_r) - \mathcal{A}_{\beta_1\boldsymbol{\theta}}(t_r)\mathcal{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(t_r)\mathcal{A}_{\boldsymbol{\theta}\beta_1}(t_r). \quad (3.29)$$

It can be shown that $M_1(t_r)/M_3^2(t_r) = \Omega_{\beta_1\beta_1}(t_r)$ where $\Omega_{\beta_1\beta_1}(t_r)$ is the diagonal entry of the asymptotic variance $\Omega(t_r)$ in (3.21) corresponding to β_1 . Based on (3.24), (3.27) and (3.28) the sequential estimators

$$\left\{ \sqrt{n} \left(\widehat{\beta}_1(t_r) - \beta_1^\circ \right), r = 1, \dots, R \right\} \quad (3.30)$$

follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\Gamma = (\gamma_{rs})$, $1 \leq r \leq s \leq R$, where

$$\gamma_{rs} = \begin{cases} \Omega_{\beta_1\beta_1}(t_r), & r = s = 1, \dots, R \\ M_2(t_r, t_s)/M_3^2(t_s), & 1 \leq r < s \leq R. \end{cases}$$

The test statistic of interest for the r th analysis is

$$T(t_r) = \frac{\sqrt{n_r}(\widehat{\beta}_1(t_r) - \beta_1^\circ)}{\sqrt{\Omega_{\beta_1\beta_1}(t_r)}}.$$

From (3.30) it can be shown that the R -dimensional joint distribution of

$$\{T(t_r), r = 1, \dots, R\} \xrightarrow{d} \text{MVN}(\mathbf{0}, \bar{\Gamma}), \quad (3.31)$$

where

$$\bar{\Gamma} = \frac{\gamma_{rs}}{\sqrt{\gamma_{rr}\gamma_{ss}}}, \quad 1 \leq r \leq s \leq R.$$

We reject H_0 at the r th interim analysis if the test statistic $T(t_r)$ exceeds the corresponding critical value. The serial critical values for the sequential test statistics can be obtained either through standard values directly if the independent increments structure (IIS) (Spiessens et al, 2000) of the statistics holds (see Section 3.7.3), or through the derivation based on the multivariate joint distribution of sequential test statistics as discussed in Section 3.10.

3.7.3 The Independent Increments Structure

The independent increments structure (IIS) refers to a property of sequential test statistics where the random variables providing data between the current analysis time at t_r and the next analysis time t_{r+1} are statistically independent of random variables providing the current data (Spiessens et al, 2000). Lee et al (1996) proved that $n^{-\frac{1}{2}}\bar{U}(t_r; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha})$ has the asymptotic independent increments structure if the covariance function $\Sigma(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha})$ is

correctly specified. Based on (3.27) the asymptotically independent increments structure of the sequentially computed test statistics can be assessed by

$$M_1(t_r) = M_2(t_r, t_s), \quad r < s \quad (3.32)$$

through

$$\text{cov}(n^{-\frac{1}{2}}\bar{U}(t_r; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha}), n^{-\frac{1}{2}}\bar{U}(t_s; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha})) = \text{var}(n^{-\frac{1}{2}}\bar{U}(t_r; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha})), \quad r < s.$$

If the independent increments property holds then the established critical values for standard group sequential boundaries of sequential test statistics can be used; these are computed through recursive univariate integration as described by [Armitage et al \(1969\)](#). If the independent increments structure does not hold, the joint distribution of the sequential test statistics, and in particular the covariance matrix of successive test statistics in (3.31) must be used to obtain critical values which enables control of the type I error rate. We show how this can be carried out in practice in Section 3.10.

3.8 Results for a Simple Two-Arm Trial

3.8.1 Derivations for Matrices $\mathcal{A}(t_r)$, $\mathcal{B}(t_r)$ and $\mathcal{B}(t_r, t_s)$

If we constrain $\beta_2 = \mathbf{0}$ (i.e. we do not adjust for covariates) and let $\boldsymbol{\beta} = (\beta_1, \beta_0)'$, then under the working independence assumption we can write the generalized estimating function in (3.18) as

$$\mathbf{U}_i(t_r; \boldsymbol{\beta}) = \Delta_i(t_r) \sum_{j=1}^{K_i(t_r)} \begin{pmatrix} U_{ij1}(\boldsymbol{\beta}) \\ U_{ij2}(\boldsymbol{\beta}) \end{pmatrix}$$

where, with the Poisson variance function, $U_{ij1}(\boldsymbol{\beta}) = X_i(Y_{ij} - \mu(X_i))$ is the elementary estimating function corresponding to β_1 and $U_{ij2}(\boldsymbol{\beta}) = Y_{ij} - \mu(X_i)$ corresponds to β_0 . The $\mathcal{A}(t_r)$ matrix in (3.22) is given more explicitly by

$$\mathcal{A}(t_r) = \pi(t_r) \begin{pmatrix} \kappa_1(t_r) \mu_1 P(X_i = 1) & \kappa_1(t_r) \mu_1 P(X_i = 1) \\ \kappa_1(t_r) \mu_1 P(X_i = 1) & \sum_{x=0}^1 \kappa_x(t_r) \mu_x P(X_i = x) \end{pmatrix},$$

where $\kappa_x(t_r) = \mathbb{E}\{K_i(t_r) \mid X_i = x, \Delta_i(t_r) = 1\}$ is the expected number of donations for a subject in group x for the r th interim analysis, $\mu_x = \mu(X_i = x; \boldsymbol{\beta})$ for $x = 0, 1$, and $\pi(t_r) = P(\Delta_i(t_r) = 1)$.

We again assume an exchangeable pairwise correlation ρ for responses within individuals, in which case

$$\text{cov}(Y_{ij}, Y_{ik} \mid X_i = x) = \rho \cdot \text{var}(Y_{ij} \mid X_i = x).$$

For the $\mathcal{B}(t_r)$ matrix in (3.20) we have

$$\mathcal{B}(t_r) = \pi(t_r) \begin{pmatrix} P(X_i = 1) \{\kappa_1(t_r)V_1 + \delta_1(t_r)\rho V_1\} & P(X_i = 1) \{\kappa_1(t_r)V_1 + \delta_1(t_r)\rho V_1\} \\ P(X_i = 1) \{\kappa_1(t_r)V_1 + \delta_1(t_r)\rho V_1\} & \sum_{x=0}^1 P(X_i = x) \{\kappa_x(t_r)V_x + \delta_x(t_r)\rho V_x\} \end{pmatrix},$$

where $\delta_x(t_r) = \nu_x(t_r) + \kappa_x^2(t_r) - \kappa_x(t_r)$, $\nu_x(t_r) = \text{var}(K_i(t_r) \mid X_i = x, \Delta_i(t_r) = 1)$ is the variance of the number of donations for a subject in group x for the r th interim analysis, and $V_x = \text{var}(Y_{ij} \mid X_i = x)$, $x = 0, 1$.

If $t_r < t_s$ then $K_i(t_r) \leq K_i(t_s)$ and the matrix $\mathcal{B}(t_r, t_s)$ in (3.26) is given by

$$\pi(t_r) \begin{pmatrix} P(X_i = 1) \{\kappa_1(t_r)V_1 + \delta_1(t_r, t_s)\rho V_1\} & P(X_i = 1) \{\kappa_1(t_r)V_1 + \delta_1(t_r, t_s)\rho V_1\} \\ P(X_i = 1) \{\kappa_1(t_r)V_1 + \delta_1(t_r, t_s)\rho V_1\} & \sum_{x=0}^1 P(X_i = x) \{\kappa_x(t_r)V_x + \delta_x(t_r, t_s)\rho V_x\} \end{pmatrix},$$

where $\delta_x(t_r, t_s) = \nu_x(t_r) + \kappa_x^2(t_r) + \kappa_x(t_r)\kappa_x(t_r, t_s) - \kappa_x(t_r)$, $\kappa_x(t_r, t_s) = \mathbb{E}\{K_i(t_r, t_s) \mid X_i = x, \Delta_i(t_r) = 1\}$ is the expected number of donations from study time t_r to t_s for a subject i in group x , $x = 0, 1$, and $K_i(t_r, t_s) = K_i(t_s) - K_i(t_r)$ denotes the number of donations over time period $(t_r, t_s]$ for individual i . See Appendix B.4 for details of the derivation.

3.8.2 Derivations for $M_1(t_r)$, $M_2(t_r, t_s)$ and $M_3(t_r)$

With equal allocation $P(X_i = 1) = P(X_i = 0) = 0.5$, the asymptotic variance of $\bar{U}(t_r; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha})/\sqrt{n}$ in (3.23) can be written as

$$M_1(t_r) = \pi(t_r) \frac{\sum_{x=0}^1 \kappa_x^2(t_r) \mu_x^2 V_{1-x} \{\kappa_{1-x}(t_r) + \delta_{1-x}(t_r)\rho\}}{2\{\kappa_0(t_r)\mu_0 + \kappa_1(t_r)\mu_1\}^2}, \quad (3.33)$$

while $\text{cov}(\bar{U}_i(t_r; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha}), \bar{U}_i(t_s; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha}))$ in (3.25) is

$$M_2(t_r, t_s) = \pi(t_r) \frac{\sum_{x=0}^1 \kappa_x(t_r) \kappa_x(t_s) \mu_x^2 V_{1-x} \{\kappa_{1-x}(t_r) + \delta_{1-x}(t_r, t_s) \rho\}}{2\{\kappa_0(t_r) \mu_0 + \kappa_1(t_r) \mu_1\} \{\kappa_0(t_s) \mu_0 + \kappa_1(t_s) \mu_1\}}. \quad (3.34)$$

Finally $M_3(t_r)$ in (3.29) can be shown to equal

$$M_3(t_r) = \pi(t_r) \frac{\kappa_0(t_r) \kappa_1(t_r) \mu_0 \mu_1}{2\{\kappa_0(t_r) \mu_0 + \kappa_1(t_r) \mu_1\}}; \quad (3.35)$$

see Appendix B.5 for details of the derivation.

3.8.3 The Asymptotic Variance of $\hat{\beta}_1(t_r)$

The inverse of $\mathcal{A}(t_r)$ is given by

$$\mathcal{A}^{-1}(t_r) = \pi(t_r)^{-1} \begin{pmatrix} \frac{2}{\kappa_0(t_r) \mu_0} + \frac{2}{\kappa_1(t_r) \mu_1} & -\frac{2}{\kappa_0(t_r) \mu_0} \\ -\frac{2}{\kappa_0(t_r) \mu_0} & \frac{2}{\kappa_0(t_r) \mu_0} \end{pmatrix},$$

which gives $\Omega(t_r) = \mathcal{A}^{-1}(t_r) \mathcal{B}(t_r) \{\mathcal{A}^{-1}(t_r)\}'$ in (3.21) as

$$\Omega(t_r) = \pi(t_r)^{-1} \begin{pmatrix} \sum_{x=0}^1 \frac{2V_x}{\kappa_x^2(t_r) \mu_x^2} [\kappa_x(t_r) + \delta_x(t_r) \rho] & -\frac{2V_0}{\kappa_0^2(t_r) \mu_0^2} [\kappa_0(t_r) + \delta_0(t_r) \rho] \\ -\frac{2V_0}{\kappa_0^2(t_r) \mu_0^2} [\kappa_0(t_r) + \delta_0(t_r) \rho] & \frac{2V_0}{\kappa_0^2(t_r) \mu_0^2} [\kappa_0(t_r) + \delta_0(t_r) \rho] \end{pmatrix}. \quad (3.36)$$

Thus the asymptotic variance of $\sqrt{n}(\hat{\beta}_1(t_r) - \beta_1^\circ)$ is the (1, 1) entry of (3.36), i.e.,

$$\Omega_{\beta_1 \beta_1}(t_r) = \frac{1}{\pi(t_r)} \left\{ \frac{2V_0}{\kappa_0(t_r) \mu_0^2} \left(1 + \frac{\delta_0(t_r) \rho}{\kappa_0(t_r)} \right) + \frac{2V_1}{\kappa_1(t_r) \mu_1^2} \left(1 + \frac{\delta_1(t_r) \rho}{\kappa_1(t_r)} \right) \right\}, \quad (3.37)$$

and the asymptotic variance of $\hat{\beta}_1(t_r)$ is obtained by dividing (3.37) by n_r . See Appendix B.5 for further details.

3.9 Modelling Accrual Time and Donation Rates

As shown in (3.33) to (3.37) in Section 3.8, the mean and variance of the number of measurements per individual are required to evaluate the joint distribution of the sequential test statistics for the sample size calculation. Suppose the number of transfusions per individual over a period of duration τ is $\mathcal{K}_i(\tau) = K$. As noted in Section 3.7.1, if the number of donations per individual varies according to a mixed time-homogeneous Poisson process, then conditionally on u , $K \mid u \sim \text{Poisson}(u\lambda\tau)$, and

$$P(K = k \mid u) = (u\lambda\tau)^k e^{-u\lambda\tau} / k! ,$$

where $\mathbb{E}(U) = 1$ and $\text{var}(U) = \phi$. We consider a negative binomial process for which $U \sim \text{Gamma}(\phi^{-1}, \phi)$. Suppose the recruitment time of a random subject in the population given its random effect follows an exponential distribution with

$$L \mid U = u \sim \text{Exp}(u\lambda) ,$$

where the random effect $U \sim \text{Gamma}(\phi^{-1}, \phi)$. We note that subjects who have higher frequency of donations in calendar time will tend to be recruited earlier in the accrual period. Thus for the trial design with interim monitoring we need to consider the recruitment time of the individual when modeling the number of donations they have by each interim analysis time. The conditional distribution of the random effect U given recruitment time $L = l$ then follows a modified gamma distribution with

$$U \mid L = l \sim \text{Gamma}\left(\phi^{-1} + 1, \frac{\phi}{\phi\lambda l + 1}\right) .$$

With $\mathbb{E}(U \mid L = l) = (1 + \phi)/(1 + \phi\lambda l)$ this reflects the fact that individuals with an earlier recruitment time tend to have larger random effects, which contributes to them providing a larger number of repeated measurements during the follow-up period compared to those recruited at a later time.

To compute the expected number of donations over the course of the trial to time t_r we average over the recruited time for recruited individuals. The mean of the number of donations per individual by study time t_r with $X_i = x$ is $\kappa_x(t_r)$ and obtained as

$$\begin{aligned} \kappa_x(t_r) &= \mathbb{E}\{K_i(t_r) \mid X_i = x, \Delta_i(t_r) = 1\} \\ &= \mathbb{E}_L\left\{\mathbb{E}_U\left\{\mathbb{E}\{K_i(t_r) \mid X_i = x, \Delta_i(t_r) = 1, \right. \right. \\ &\quad \left. \left. U_i = u, L_i = l\} \mid \Delta_i(t_r) = 1, L_i < \min(t_r, \tau_A)\right\} \mid \Delta_i(t_r) = 1, L_i < \min(t_r, \tau_A)\right\} . \end{aligned}$$

For $0 < t_r \leq \tau$ we have

$$\kappa_x(t_r) = \begin{cases} \lambda \mathbb{E} \{u \mid l < \min(t_r, \tau_A)\} [t_r - \mathbb{E}\{l \mid l < \min(t_r, \tau_A)\}] , & 0 < t_r \leq \tau_F \\ \lambda \mathbb{E} \{u \mid l < \min(t_r, \tau_A)\} \mathbb{E} \{\min(t_r - l, \tau_F) \mid l < \min(t_r, \tau_A)\} , & \tau_F < t_r \leq \tau \end{cases}$$

where

$$\mathbb{E}\{u \mid l < t\} = \frac{1 - (\lambda\phi t + 1)^{-\phi^{-1}-1}}{1 - (\lambda\phi t + 1)^{-\phi^{-1}}} .$$

Similarly, $\nu_x(t_r)$, the variance of the number of donations per individual by study time t_r and the mean number of donations per individual between study time t_r and t_s (i.e. $\kappa_x(t_r, t_s)$) where $0 < t_r < t_s \leq \tau$, can be calculated via

$$\nu_x(t_r) = \begin{cases} \kappa_x(t_r) - \kappa_x^2(t_r) + \lambda^2 \mathbb{E} \{u^2 \mid l < \min(t_r, \tau_A)\} \times \\ \quad [t_r^2 + \mathbb{E}\{l^2 \mid l < \min(t_r, \tau_A)\} - 2t_r \mathbb{E}\{l \mid l < \min(t_r, \tau_A)\}] , & 0 < t_r \leq \tau_F \\ \kappa_x(t_r) - \kappa_x^2(t_r) + \lambda^2 \mathbb{E} \{u^2 \mid l < \min(t_r, \tau_A)\} \times \\ \quad \mathbb{E} \{(\min(t_r - l, \tau_F))^2 \mid l < \min(t_r, \tau_A)\} , & \tau_F < t_r \leq \tau \end{cases}$$

and

$$\kappa_x(t_r, t_s) = \begin{cases} \lambda(t_s - t_r) \mathbb{E} \{u \mid l < \min(t_r, \tau_A)\} , & 0 < t_r < t_s \leq \tau_F \\ \lambda \mathbb{E} \{\max(\min(t_s - t_r, l + \tau_F - t_r), 0) \mid l < \min(t_r, \tau_A)\} \times \\ \quad \mathbb{E} \{u \mid l < \min(t_r, \tau_A)\} , & \text{o.w.} , \end{cases}$$

where

$$\mathbb{E}\{u^2 \mid L < t\} = (1 + \phi) \frac{1 - (\lambda\phi t + 1)^{-\phi^{-1}-2}}{1 - (\lambda\phi t + 1)^{-\phi^{-1}}} .$$

See Appendix B.6 for details of derivations and explicit forms of $\kappa_x(t_r)$, $\nu_x(t_r)$, and $\kappa_x(t_r, t_s)$. Since the number of donations per recruited individual does not depend on the treatment arms here we have $\kappa(t_r) = \kappa_x(t_r)$, $\nu(t_r) = \nu_x(t_r)$, and $\kappa(t_r, t_s) = \kappa_x(t_r, t_s)$, $x = 0, 1$. With $\kappa(t_r)$, $\nu(t_r)$, and $\kappa(t_r, t_s)$ derived we can evaluate the joint distribution of sequential test statistics which is the key for the calculations of the group sequential boundaries and sample size for interim monitoring.

For a study without interim analysis, let K_A denote the number of donations from a random subject during the accrual period of length τ_A , and K_F be the number of donations from the subject in the follow-up window of length τ_F . Recall that in Section 3.3.2 we show that the mean of K_F given that subject has at least one donation during the recruitment phase ($K_A \geq 1$) is

$$\mathbb{E}(K_F \mid K_A \geq 1) = \lambda\tau_F \left\{ \frac{1 - (1 + \lambda\tau_A\phi)^{-\phi^{-1}-1}}{1 - (1 + \lambda\tau_A\phi)^{-\phi^{-1}}} \right\},$$

and the variance of K_F given $K_A \geq 1$ is

$$\text{var}(K_F \mid K_A \geq 1) = \mathbb{E}(K_F \mid K_A \geq 1) - \mathbb{E}^2(K_F \mid K_A \geq 1) + (\lambda\tau_F)^2(1 + \phi) \left\{ \frac{1 - (1 + \lambda\tau_A\phi)^{-\phi^{-1}-2}}{1 - (1 + \lambda\tau_A\phi)^{-\phi^{-1}}} \right\}.$$

It can be shown that $\kappa(\tau) = \mathbb{E}(K_F \mid K_A \geq 1)$ and $\nu(\tau) = \text{var}(K_F \mid K_A \geq 1)$. Hence the modeling of the recruitment time on number of donations is a special case for the design of sequential analysis.

3.10 Sample Size and Group Sequential Testing

3.10.1 Alpha Spending Functions

Let $C(t_r)$ be the critical value or group sequential boundary at t_r such that if $|T(t_r)| > C(t_r)$ we reject H_0 and may consider the termination of the study. Let $\alpha^*(t)$ be an increasing function with $\alpha^*(0) = 0$ and $\alpha^*(1) = \alpha$ where α is the predetermined significance level. The alpha spending function $\alpha^*(t)$ proposed by [Lan and DeMets \(1983\)](#) allocated the significance level α over the interim analyses, and based on the alpha spending function the group sequential boundaries can be determined by solving

$$\begin{aligned} P(|T(t_1)| > C(t_1) \mid \beta_1 = \beta_1^\circ) + P(|T(t_1)| \leq C(t_1), |T(t_2)| > C(t_2) \mid \beta_1 = \beta_1^\circ) + \dots \\ + P(|T(t_r)| \leq C(t_r), r = 1, \dots, R-1, |T(t_R)| > C(t_R) \mid \beta_1 = \beta_1^\circ) = \alpha^*(\omega_R), \end{aligned}$$

or

$$\begin{aligned} P(|T(t_1)| > C(t_1) \mid \beta_1 = \beta_1^\circ) &= \alpha^*(\omega_1), \text{ and} \\ P(|T(t_v)| \leq C(t_v), v = 1, \dots, r-1, |T(t_r)| > C(t_r) \mid \beta_1 = \beta_1^\circ) &= \alpha^*(\omega_r) - \alpha^*(\omega_{r-1}), \\ & r = 2, \dots, R, \end{aligned} \quad (3.38)$$

where ω_r is the information fraction for the r th interim analysis, $0 \leq \omega_r \leq 1$. There are three common ways to define the information fraction. The first is using the statistical information which is the inverse of the variance of the estimated effect (e.g. $n_r/\Omega_{\beta_1\beta_1}(t_r)$ for the r th interim analysis). If the maximum information is known or can be specified at the design stage based on historical research, the information fraction can be defined as $\omega_r = I(t_r)/I(\tau)$ where $I(t_r) = n_r/\Omega_{\beta_1\beta_1}(t_r)$. This can also be used as the guidance of determining the time of interim analysis. If the maximum information is unknown we can instead use the second approach based on the expected number of subjects/measurements (i.e. $\omega_r = n_r/n$), The third is to specify it in terms of the expected study duration (i.e. $\omega_r = t_r/\tau$).

[Lan and DeMets \(1983\)](#) proposed the use of alpha spending functions for which the number of interim analyses is not required in advance for the calculation of critical values. Examples include

$$\begin{aligned} \alpha_1^*(t) &= 2 - 2\Phi(Z_{\alpha/2}/\sqrt{t}), \\ \alpha_2^*(t) &= \alpha \log\{1 + (e-1)t\}, \\ \alpha_3^*(t) &= \alpha t. \end{aligned}$$

The boundaries generated by function $\alpha_1^*(t)$ and $\alpha_2^*(t)$ are similar to the boundaries obtained by [O'Brien and Fleming \(1979\)](#) and [Pocock \(1977\)](#) respectively, and we refer them to the O'Brien-Fleming and Pocock type alpha spending functions respectively. The third one $\alpha_3^*(t)$ is the uniform alpha spending function as it allocates the type I error rate evenly over time ([Lan and DeMets, 1983](#)).

3.10.2 Sample Size for a Sequential Superiority Trial

For a superiority trial we consider a two-sided test with the aim to test

$$H_0 : \beta_1 = \beta_1^\circ = 0 \text{ versus } H_A : \beta_1 \neq 0$$

where the true effect is denoted as β_{1A} . Using the joint distribution of the sequential test statistics in (3.31) and assumptions about the response process, accrual, and transfusions discussed in Section 3.7.1 and Section 3.9, it is possible to determine the times of interim analyses that will give a dependence structure aligned with the independent increments property by verifying (3.32). This will enable the use of established critical values directly (Lee et al, 1996). The times of interim analysis can also be determined based on the proportions of the maximum information, e.g. times corresponding to 1/3, 2/3 of the maximum information for a study with two interim analyses and one final analysis. In such cases or more generally for analyses planned at any key times of interest, the critical values can be derived through (3.38) to ensure the control of the type I error rate at α for the specified alpha spending function. The sample size for a sequential superiority study can be calculated by solving (3.39) with the desired power,

$$\text{Power} = P(|T(t_1)| > C(t_1) \mid \beta_1 = \beta_{1A}) + P(|T(t_1)| \leq C(t_1), |T(t_2)| > C(t_2) \mid \beta_1 = \beta_{1A}) + \dots + P(|T(t_r)| \leq C(t_r), r = 1, \dots, R-1, |T(t_R)| > C(t_R) \mid \beta_1 = \beta_{1A}). \quad (3.39)$$

Note that the number of interim analyses is not required for the calculation of critical values based on Lan and DeMets (1983) and we determine the number of interim analyses here for study design purpose only.

3.10.3 Sample Size for a Sequential Non-Inferiority Trial

For a non-inferiority design we consider a one-sided test with the aim to test

$$H_0 : \beta_1 \geq \beta_{1M} \text{ versus } H_A : \beta_1 < \beta_{1M}$$

where β_{1M} is the log-relative rate at the margin value of β_1 , i.e. $\beta_{1M} = \log RR_M$, and for $RR < RR_M$ we can claim non-inferiority on the relative rate scale. This corresponds to the case where the relative rate of an adverse event for the blood donation using the experimental system versus the standard system is no greater than RR_M . Based on (3.38) the critical values for the sequential tests for the non-inferiority trial can be solved through

$$P(T(t_1) < -C(t_1) \mid \beta_1 = \beta_1^\circ) = \alpha^*(\omega_1), \quad r = 1, \text{ and} \\ P(-C(t_v) \leq T(t_v) \leq 0, v = 1, \dots, r-1, T(t_r) < -C(t_r) \mid \beta_1 = \beta_1^\circ) = \alpha^*(\omega_r) - \alpha^*(\omega_{r-1}), \\ r = 2, \dots, R.$$

Once the critical values have been determined, the sample size for the study with interim analyses can be calculated by solving (3.40) such that the desired power requirement is met,

$$\begin{aligned} \text{Power} = & P(T(t_1) < -C(t_1) \mid \beta_1 = \beta_{1A}) + \\ & P(-C(t_1) \leq T(t_1) \leq 0, T(t_2) < -C(t_2) \mid \beta_1 = \beta_{1A}) + \\ & \cdots + P(-C(t_r) \leq T(t_r) \leq 0, r = 1, \dots, R-1, T(t_R) < -C(t_R) \mid \beta_1 = \beta_{1A}). \end{aligned} \quad (3.40)$$

3.11 Simulation Studies

3.11.1 Empirical Validation with Sequential Trials

In this section we report on the results of simulation studies designed to validate the derivations. Let $\lambda = 10, 18$, $\phi = 0.8, 1.6$, $\tau_A = 0.5$, $\tau_F = 1$, $\rho = 0.005, 0.01$, $\mu_0 = 0.002$, $\beta_{1A} = \log(1.2)$, and type I error rates are controlled at the significance level of $\alpha = 5\%$. Also let $\beta_1^\circ = 0$ for the superiority study, and $\beta_{1M} = \log(1.5)$ for the non-inferiority study. The sample sizes are then calculated to ensure adequate power of trials (at least 80% or 90%). Suppose we plan for two interim analyses with a final analysis at the end of the study so $R = 3$. The times of two interim analyses are determined based on the proportions of the maximum information where the first and second analysis are at the times corresponding to 1/3 and 2/3 of the maximum information, and the times of interim analyses can be determined without knowing the actual maximum information but the asymptotic variance of β_1 , i.e. $\Omega_{\beta_1\beta_1}(t_r)/n_r$, and the probability of being recruited by time t_r , i.e. $\pi(t_r)$. In this example suppose the information for the two interim analyses is the inverse of the asymptotic variance of β_1 , i.e.,

$$\mathcal{I}(t_1) = \frac{n_1}{\Omega_{\beta_1\beta_1}(t_1)} = \frac{\pi(t_1)n}{\Omega_{\beta_1\beta_1}(t_1)} \quad \text{and} \quad \mathcal{I}(t_2) = \frac{n_2}{\Omega_{\beta_1\beta_1}(t_2)} = \frac{\pi(t_2)n}{\Omega_{\beta_1\beta_1}(t_2)},$$

and the maximum information at the end of the study is $\mathcal{I}(\tau) = n/\Omega_{\beta_1\beta_1}(\tau)$. By taking the ratio of the information at t_1 and t_2 with the maximum information respectively gives two information fractions (i.e., $\mathcal{I}(t_1)/\mathcal{I}(\tau)$ and $\mathcal{I}(t_2)/\mathcal{I}(\tau)$), and the times of two interim analyses are chosen at the times such that the information fractions equal to 1/3 and 2/3 respectively, i.e., t_1 and t_2 are determined such that

$$\frac{\mathcal{I}(t_1)}{\mathcal{I}(\tau)} = \frac{\pi(t_1)\Omega_{\beta_1\beta_1}(\tau)}{\Omega_{\beta_1\beta_1}(t_1)} = \frac{1}{3} \quad \text{and} \quad \frac{\mathcal{I}(t_2)}{\mathcal{I}(\tau)} = \frac{\pi(t_2)\Omega_{\beta_1\beta_1}(\tau)}{\Omega_{\beta_1\beta_1}(t_2)} = \frac{2}{3},$$

where

$$\pi(t_r) = \frac{1 - (\lambda\phi t_r + 1)^{-\phi^{-1}}}{1 - (\lambda\phi\tau_A + 1)^{-\phi^{-1}}}.$$

Once the interim analyses times are determined, the group sequential boundaries can be derived and sample sizes needed for the study with two interim analyses can be calculated as described in Section 3.10.2 and 3.10.3. Three different alpha spending functions are considered. The required sample sizes for each scenario are provided in Table 3.3 for the superiority study design, and Table 3.4 for the non-inferiority study. The “Fixed study” columns of Table 3.3 and Table 3.4 record the sample sizes required for a fixed study without any interim analysis. The columns after “Fixed study” record the sample sizes required for the study with interim monitoring using different alpha spending functions. The sample size needed for a study with interim analyses increases compared to a fixed sample size study in order to reach the same desired power since the net threshold for significance is higher.

Table 3.3: Sample size required in terms of the number of subjects needed for a superiority design with $\mu_0 = 0.002$, $\exp(\beta_{1A}) = 1.2$, $\tau_A = 0.5$, $\tau_F = 1$, and $\beta_1^\circ = 0$; two interim analyses are planned with $R = 3$.

					80% Power				90% Power			
λ	ϕ	ρ	κ	ν	Fixed study	Interim monitoring with alpha spending function			Fixed study	Interim monitoring with alpha spending function		
						$\alpha_1^*(t)$	$\alpha_2^*(t)$	$\alpha_3^*(t)$		$\alpha_1^*(t)$	$\alpha_2^*(t)$	$\alpha_3^*(t)$
10	0.80	0.005	12.24	91.69	38610	39530	45241	43199	51687	52823	59734	57276
10	0.80	0.010	12.24	91.69	41915	42811	49014	46790	56113	57219	64710	62033
10	1.60	0.005	14.01	190.72	34923	35755	40919	39073	46752	47782	54027	51803
10	1.60	0.010	14.01	190.72	39026	39863	45633	43567	52244	53280	60244	57758
18	0.80	0.005	20.23	277.35	24872	25421	29107	27788	33296	33981	38427	36840
18	0.80	0.010	20.23	277.35	28389	28965	33155	31651	38004	38716	43770	41961
18	1.60	0.005	22.72	577.77	23493	24009	27482	26237	31450	32087	36283	34783
18	1.60	0.010	22.72	577.77	27975	28526	32636	31157	37450	38129	43080	41306

With the calculated sample size for study with the planned interim analyses for each scenario we assess the empirical rejection rates to verify the sample size calculation. Under each parameter setting, for each subject i we generate random effect $u_i \sim \text{Gamma}(\phi^{-1}, \phi)$ and treatment assignment $x_i \sim \text{Bernoulli}(0.5)$. The recruitment time for subject i is generated from $l_i \mid u_i \sim \text{Exp}(u_i\lambda)$. We repeat the above steps and keep the subjects with

Table 3.4: Sample size required in terms of the number of subjects needed for a non-inferiority design with $\mu_0 = 0.002$, $\exp(\beta_{1A}) = 1.2$, $\tau_A = 0.5$, $\tau_F = 1$, and non-inferiority margin $\exp(\beta_{1M}) = 1.5$; two interim analyses are planned with $R = 3$.

λ	ϕ	ρ	κ	ν	80% Power				90% Power			
					Fixed study	Interim monitoring with alpha spending function			Fixed study	Interim monitoring with alpha spending function		
						$\alpha_1^*(t)$	$\alpha_2^*(t)$	$\alpha_3^*(t)$		$\alpha_1^*(t)$	$\alpha_2^*(t)$	$\alpha_3^*(t)$
10	0.80	0.005	12.24	91.69	20303	20748	23913	22758	28123	28692	32640	31206
10	0.80	0.010	12.24	91.69	22042	22473	25918	24658	30531	31079	35373	33808
10	1.60	0.005	14.01	190.72	18365	18766	21628	20584	25438	25951	29522	28225
10	1.60	0.010	14.01	190.72	20522	20927	24128	22956	28426	28944	32926	31475
18	0.80	0.005	20.23	277.35	13079	13345	15390	14642	18117	18456	21004	20077
18	0.80	0.010	20.23	277.35	14929	15205	17535	16681	20678	21029	23928	22869
18	1.60	0.005	22.72	577.77	12354	12602	14530	13825	17112	17429	19831	18955
18	1.60	0.010	22.72	577.77	14711	14976	17257	16418	20377	20710	23549	22509

$l_i \leq \tau_A$ until we get the calculated sample size n . Next we generate the waiting times between measurements for each subject according to $\text{Exp}(u_i \lambda)$. The number of donations for each subject is then determined by the accumulated waiting time for the donations within τ_F . Note that the first donation of each subject is assumed to occur at the entry time l_i of the study so the number of donations will be added by one for each subject. The correlated binary response of subject i for the k th measurement Y_{ik} is generated given the pre-specified marginal means and pairwise associations based on the method of [Qaqish \(2003\)](#). The response Y_{i1} is generated from Bernoulli distribution with mean $P(Y_{ik} = 1 \mid \mathbf{Z}_{i1})$, and response Y_{ik} , $k = 2, 3, \dots, K_i(\tau)$, can be generated sequentially though

$$\mathbb{E}(Y_{ik} \mid \bar{\mathbf{Y}}_{ik}, \bar{\mathbf{Z}}_{ik}, \mathbf{Z}_{ik}) = \mathbb{E}(Y_{ik} \mid \mathbf{Z}_{ik}) + \sum_{j=1}^{k-1} b_{ikj}(y_{ij} - \mathbb{E}(Y_{ij} \mid \mathbf{Z}_{ij})), \quad k = 2, 3, \dots, K_i(\tau),$$

where $\bar{\mathbf{Y}}_{ik} = (Y_{i1}, \dots, Y_{i,k-1})'$, $\bar{\mathbf{Z}}_{ik} = (\mathbf{Z}_{i1}, \dots, \mathbf{Z}_{i,k-1})'$, and b_{ikj} is the j th element of

$$\mathbf{b}_{ik}(\bar{\mathbf{Z}}_{ik}, \mathbf{Z}_{ik}) = \text{cov}(\bar{\mathbf{Y}}_{ik} \mid \bar{\mathbf{Z}}_{ik})^{-1} \text{cov}(\bar{\mathbf{Y}}_{ik}, Y_{ik} \mid \bar{\mathbf{Z}}_{ik}, \mathbf{Z}_{ik}), \quad k = 2, 3, \dots, K_i(\tau).$$

Following the above simulation procedure we have the recruitment times, repeated mea-

surement times and responses generated for all individuals in the sample. We next plan 2 interim analyses at t_1 and t_2 with a final analysis at the end of study. The empirical rejection rates under H_0 and H_A respectively based on the simulated data using the computed critical values are calculated and repeated for $n_{sim} = 1000$ times. The averaged empirical rejection rates for each scenario are summarized in Table 3.5 for the superiority study and in Table 3.6 for the non-inferiority study. From the summary tables we see that the empirical type I errors (empirical rejection rate under H_0) for most of the scenarios are within 5%. Similarly for the empirical power (empirical rejection rate under H_A), the empirical powers reach the desired power at either 80% or 90% settings for most of the scenarios, and some cases have empirical powers slightly below the target power due to randomness. The overall empirical rejection rates are satisfying which verify our sample size calculation method for the study with interim analyses for the cluster data.

3.11.2 Design of a Plasma Donation Study

To design the study based on the aims of the IMPACT trial as introduced in Section 1.3.2 and Section 3.4.3, we assume $P(X_i = 1) = 0.5$ so that donors are randomized into two groups with a 50:50 allocation to either the PPN or the standard nomogram arm. Based on Canadian plasma donor data on the expected annual number of donations per donor is approximately 21 so we set $\lambda = 21$. To account for variability in donation frequency among donors we set $\phi = 0.80$. The rate of significant hypotensive adverse events in the control arm using the standard nomogram is set to $\mu_0 = 0.035\%$. For the experimental arm we anticipate a 50% increase in the event rate, so we let $\exp(\beta_{1A}) = 1.50$. Given the low event rate we set a non-inferiority margin of $\exp(\beta_{1M}) = 2.5$, and consider trial planning with $H_0 : \beta_1 \geq \beta_{1M}$ versus $H_A : \beta_1 < \beta_{1M}$ with significant level $\alpha = 5\%$ and target power at 80%. We assume an exchangeable within-donor serial correlation of $\rho = 0.005$ across repeated donations. Furthermore we plan for a 6-month recruitment duration ($\tau_A = 0.5$) and follow each donor for 6 months ($\tau_F = 0.5$).

We plan to conduct two interim analyses at times t_1 and t_2 , along with a final analysis at the end of the study, $t_3 = \tau_A + \tau_F = 1$. The timing of the interim analyses can be determined by the amount of information accumulated up to each point in the study. Figure 3.5(a) illustrates the growth of the information fraction over the course of the study based on the inverse variance of the estimator $\hat{\beta}_1$. For example, interim analyses can be conducted at study times approximately $t_1 = 0.20$ and $t_2 = 0.40$ which corresponds to 1/3 and 2/3 of the maximum information expected by the end of the study. If the estimated variance of $\hat{\beta}_1$ is unavailable, alternative metrics such as the proportion of donors recruited, donations accumulated, or simply the passage of study time can be used to approximate

Table 3.5: Empirical rejection rates for a superiority trial with two interim analyses.

λ	ϕ	ρ	t	$\alpha_1^*(t)$				$\alpha_2^*(t)$				$\alpha_3^*(t)$			
				80% Power		90% Power		80% Power		90% Power		80% Power		90% Power	
				Emp. type I error	Emp. power	Emp. type I error	Emp. power	Emp. type I error	Emp. power	Emp. type I error	Emp. power	Emp. type I error	Emp. power	Emp. type I error	Emp. power
10	0.8	0.005	t_1	0.0	8.4	0.2	12.3	3.0	36.5	1.9	45.8	1.5	30.7	1.6	43.0
			t_2	1.9	42.9	1.9	56.0	2.0	24.8	1.2	31.6	1.5	28.3	1.1	32.9
			t_3	3.3	29.6	3.3	22.7	0.7	16.3	0.7	11.8	1.2	21.1	1.8	13.7
			Total	5.2	80.9	5.4	91.0	5.7	77.6	3.8	89.2	4.2	80.1	4.5	89.6
10	0.8	0.01	t_1	0.1	3.4	0.0	7.1	2.6	31.8	1.9	41.8	0.9	30.1	1.9	36.1
			t_2	2.1	44.5	1.2	56.0	1.5	29.1	1.7	34.2	1.6	29.4	2.8	36.2
			t_3	4.0	32.5	3.4	26.8	1.2	16.5	1.3	14.7	1.5	21.1	0.9	17.6
			Total	6.2	80.4	4.6	89.9	5.3	77.4	4.9	90.7	4.0	80.6	5.6	89.9
10	1.6	0.005	t_1	0.1	8.1	0.3	11.3	2.0	37.6	2.7	50.3	1.6	32.7	1.5	41.3
			t_2	2.2	42.1	1.7	54.2	2.1	27.3	1.5	26.6	1.7	28.8	1.0	35.5
			t_3	1.8	28.0	2.9	23.6	1.5	14.9	1.1	13.6	2.0	17.5	2.0	14.5
			Total	4.1	78.2	4.9	89.1	5.6	79.8	5.3	90.5	5.3	79.0	4.5	91.3
10	1.6	0.01	t_1	0.1	4.3	0.2	9.1	2.2	36.6	2.4	49.3	1.7	23.0	1.4	39.5
			t_2	1.3	47.4	1.8	51.8	2.5	28.4	1.6	27.5	1.8	32.8	2.0	34.7
			t_3	3.9	29.3	2.6	28.9	0.9	16.8	1.8	13.3	1.9	20.7	1.2	14.5
			Total	5.3	81.0	4.6	89.8	5.6	81.8	5.8	90.1	5.4	76.5	4.6	88.7
18	0.8	0.005	t_1	0.1	5.1	0.0	7.8	1.6	34.0	2.5	45.0	1.8	26.7	1.4	37.8
			t_2	0.9	45.0	1.0	56.2	1.7	30.0	1.8	31.4	1.8	34.2	1.7	35.6
			t_3	3.0	30.6	2.6	25.6	1.2	17.6	0.5	13.0	1.7	18.8	0.9	16.5
			Total	4.0	80.7	3.6	89.6	4.5	81.6	4.8	89.4	5.3	79.7	4.0	89.9
18	0.8	0.01	t_1	0.0	2.1	0.0	6.4	2.2	29.6	2.7	40.5	1.6	21.8	1.3	31.6
			t_2	1.2	42.9	1.9	58.0	1.9	29.7	2.1	33.4	1.2	33.0	1.6	40.8
			t_3	3.7	34.4	2.9	28.0	0.9	20.0	1.6	15.5	1.3	23.3	1.1	17.3
			Total	4.9	79.4	4.8	92.4	5.0	79.3	6.4	89.4	4.1	78.1	4.0	89.7
18	1.6	0.005	t_1	0.0	3.8	0.0	9.1	1.8	37.3	2.5	47.0	1.6	29.4	1.1	40.8
			t_2	1.8	47.1	1.5	57.5	1.3	25.9	1.9	29.5	1.7	32.4	1.7	37.3
			t_3	2.6	29.0	3.2	23.6	0.8	17.0	0.9	12.3	1.8	19.4	1.4	13.1
			Total	4.4	79.9	4.7	90.2	3.9	80.2	5.3	88.8	5.1	81.2	4.2	91.2
18	1.6	0.01	t_1	0.0	3.1	0.1	5.7	1.7	33.9	3.4	45.3	1.2	29.3	1.4	36.3
			t_2	2.4	48.6	1.3	58.5	1.1	30.3	1.7	31.7	1.6	33.9	1.7	36.6
			t_3	3.9	30.5	3.5	27.3	0.9	16.2	0.7	13.7	1.3	17.2	0.9	16.4
			Total	6.3	82.2	4.9	91.5	3.7	80.4	5.8	90.7	4.1	80.4	4.0	89.3

Table 3.6: Empirical rejection rates for a non-inferiority trial with two interim analyses.

λ	ϕ	ρ	t	$\alpha_1^*(t)$				$\alpha_2^*(t)$				$\alpha_3^*(t)$			
				80% Power		90% Power		80% Power		90% Power		80% Power		90% Power	
				Emp. type I error	Emp. power	Emp. type I error	Emp. power	Emp. type I error	Emp. power	Emp. type I error	Emp. power	Emp. type I error	Emp. power	Emp. type I error	Emp. power
10	0.8	0.005	t_1	0.0	6.4	0.0	10.7	2.8	40.9	2.0	49.6	1.1	33.7	1.5	43.8
			t_2	2.6	45.4	1.7	52.0	1.5	26.9	2.1	27.4	1.6	30.4	1.2	31.4
			t_3	3.1	27.4	2.4	25.5	0.9	14.0	1.1	12.4	1.6	16.8	1.8	14.9
			Total	5.7	79.2	4.1	88.2	5.2	81.8	5.2	89.4	4.3	80.9	4.5	90.1
10	0.8	0.01	t_1	0.2	3.6	0.0	8.0	1.2	32.7	1.6	43.4	1.8	30.3	1.4	40.6
			t_2	2.3	41.4	1.5	52.8	1.2	29.5	1.5	29.6	1.7	32.2	2.1	32.3
			t_3	2.7	33.9	4.0	28.6	1.0	15.9	1.6	14.7	1.8	17.0	1.5	14.9
			Total	5.2	78.9	5.5	89.4	3.4	78.1	4.7	87.7	5.3	79.5	5.0	87.8
10	1.6	0.005	t_1	0.0	7.5	0.1	10.5	3.2	37.6	1.0	51.7	1.0	34.6	1.7	42.9
			t_2	2.5	43.8	2.1	55.5	0.7	25.9	0.9	27.8	1.7	27.9	1.5	33.4
			t_3	3.9	27.6	2.3	23.1	0.6	15.9	0.9	9.9	2.3	18.3	1.3	14.3
			Total	6.4	78.9	4.5	89.1	4.5	79.4	2.8	89.4	5.0	80.8	4.5	90.6
10	1.6	0.01	t_1	0.1	4.3	0.0	9.2	3.0	35.2	2.0	47.6	2.0	30.2	1.2	42.3
			t_2	1.7	45.0	2.4	56.2	1.6	30.5	2.1	29.0	1.3	30.3	1.7	35.5
			t_3	2.2	31.9	3.4	23.7	1.2	15.3	0.9	10.6	0.7	18.6	1.7	12.7
			Total	4.0	81.2	5.8	89.1	5.8	81.0	5.0	87.2	4.0	79.1	4.6	90.5
18	0.8	0.005	t_1	0.2	4.2	0.1	7.7	1.5	37.2	2.4	47.1	1.5	30.6	1.7	43.7
			t_2	1.5	45.5	1.4	55.4	1.4	25.5	1.8	31.0	1.9	29.0	1.0	32.2
			t_3	3.5	32.6	2.4	28.2	1.4	16.2	1.1	13.0	1.3	19.3	1.5	14.7
			Total	5.2	82.3	3.9	91.3	4.3	78.9	5.3	91.1	4.7	78.9	4.2	90.6
18	0.8	0.01	t_1	0.1	3.9	0.0	4.5	2.6	33.4	2.7	44.2	2.4	24.7	1.5	36.2
			t_2	1.0	42.6	1.8	53.7	1.7	27.1	1.5	31.7	1.8	32.4	1.5	38.6
			t_3	4.2	33.0	4.1	30.0	1.2	18.4	1.0	13.9	1.6	20.4	0.8	15.5
			Total	5.3	79.5	5.9	88.2	5.5	78.9	5.2	89.8	5.8	77.6	3.8	90.3
18	1.6	0.005	t_1	0.0	5.7	0.1	7.3	2.1	38.1	3.9	49.0	2.0	30.3	1.2	41.7
			t_2	1.6	43.4	2.2	56.4	1.5	26.7	1.7	28.8	2.0	30.4	2.1	35.2
			t_3	2.9	30.8	3.0	24.8	1.6	14.1	0.8	11.1	1.3	18.9	1.0	13.1
			Total	4.5	79.9	5.3	88.5	5.2	78.9	6.4	88.9	5.3	79.6	4.3	90.0
18	1.6	0.01	t_1	0.0	2.9	0.1	6.7	1.6	33.4	2.2	46.4	1.7	28.8	1.7	42.2
			t_2	1.1	48.2	1.5	56.6	2.6	30.5	1.7	33.0	1.5	31.3	1.4	31.3
			t_3	4.4	30.7	4.1	27.1	1.2	15.7	0.8	11.6	0.7	17.7	1.8	15.1
			Total	5.5	81.8	5.7	90.4	5.4	79.6	4.7	91.0	3.9	77.8	4.9	88.6

the information fraction and guide the timing of interim analyses. The critical values for three analyses, i.e. first interim analysis at t_1 , second interim analysis at t_2 , and final analysis at $t_3 = 1$, are provided in Table B.1 in Appendix B.7. These critical values are calculated based on the information fraction defined as the fraction of inverse variance of $\hat{\beta}_1$. Using an O'Brien & Fleming alpha spending function $\alpha_1^*(t)$, the critical values at $t_1 = 0.2$, $t_2 = 0.4$, and $t_3 = 1$ are 3.25, 2.09, and 1.70 respectively. For a Pocock-type spending function $\alpha_2^*(t)$, the corresponding critical values are 2.01, 1.96, and 1.98, while for a uniform spending function $\alpha_3^*(t)$ they are 2.14, 1.96, and 1.88; see Table B.1 in Appendix B.7. The O'Brien-Fleming alpha spending function heavily protects against early stopping by setting a very strict (high) critical value for the first interim analysis with more relaxed critical values toward the final analysis. Pocock type alpha spending function provides more balanced critical values across all interim analyses, potentially allowing for earlier termination of the trial since the first interim analysis has a considerably lower critical value compared to O'Brien-Fleming type. The uniform alpha spending function allocates the type I error rate uniformly across all interim analyses and it falls between O'Brien-Fleming and Pocock in terms of conservatism on early stopping.

Figure 3.5(b) shows the number of donors recruited over time. Under the same setting a non-sequential study requires the fewest donors ($n = 20,426$) to achieve 80% power and a study planned using Pocock type alpha spending function requires the largest number of donors ($n = 24,031$). If the study is terminated early following interim analyses, the total number of recruited donors will be fewer than the sample size originally planned for the end of the study. The average number of donations per recruited donor (Figure 3.5(c)) and number of donations (Figure 3.5(d)) increase rapidly during the recruitment period as the frequent donors are encountered and recruited earlier.

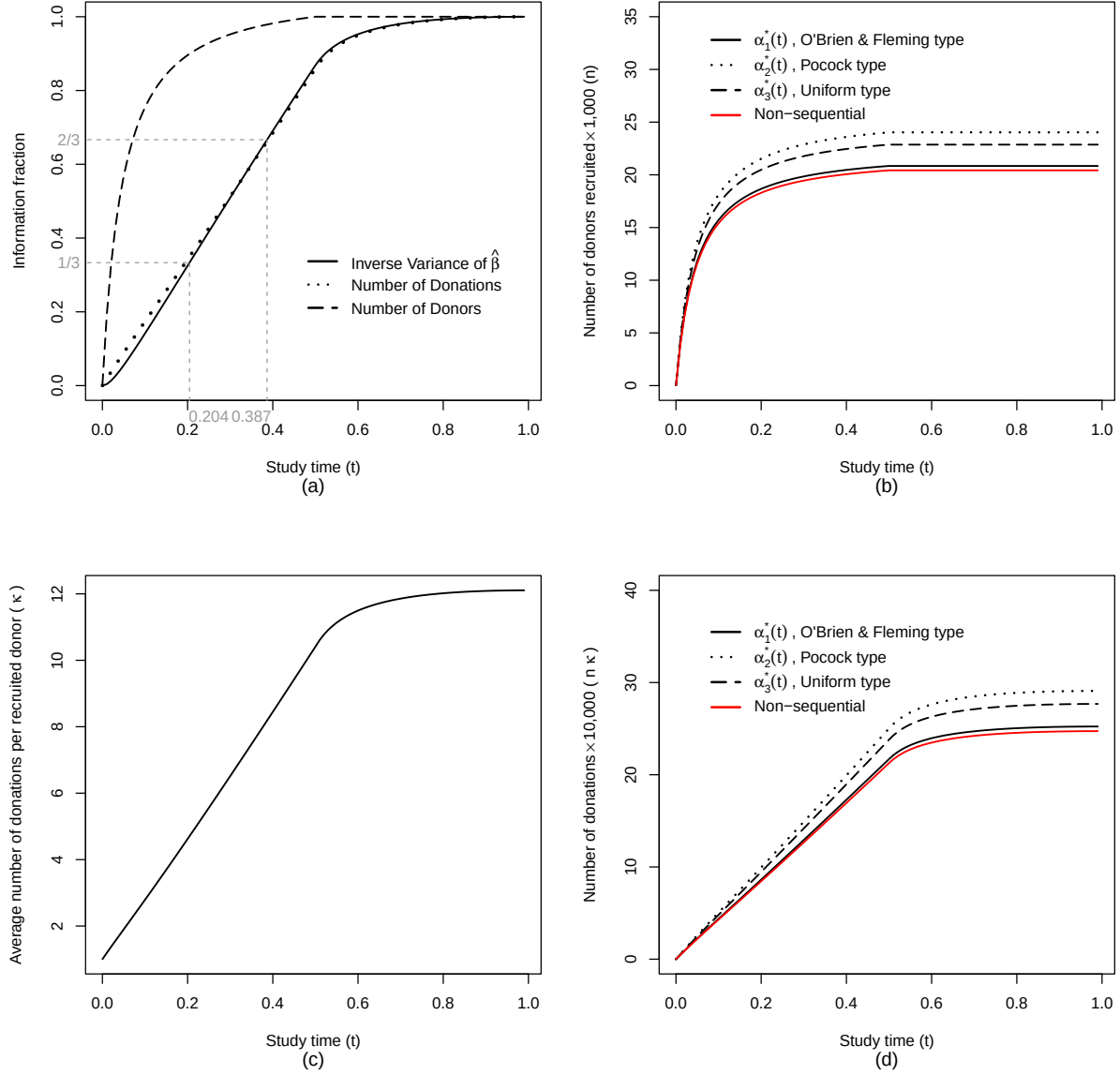


Figure 3.5: Study progression over time in a non-inferiority trial of information fraction, number of donors recruited, average donations per donor, and total donations with $\lambda = 21$, $\phi = 0.8$, $\mu_0 = 0.035\%$, $\exp(\beta_{1A}) = 1.5$, $\exp(\beta_{1M}) = 2.5$, $\rho = 0.005$, $\tau_F = 0.5$, $\alpha = 5\%$, and target power=80%.

3.12 Remarks

Many trials involve the collection of repeated observations from recruited individuals but most often the number of planned measurements is common and set by the design. Study withdrawal may mean that the actual number of measurements each individual contributes varies resulting in some individuals giving fewer than the intended measurements. There is a consequent loss in power and there has been much work aiming to mitigate biases arising from selective attrition including methods based on inverse probability weighting ([Rombach et al, 2018](#)) and multiple imputation ([Cro et al, 2020](#)). In the present setting the variation in the number of measurements is a consequence of individual-specific propensities to donate blood and there is considerable variation in the donor population in this rate. We have assessed the validity of the negative binomial model in some donor populations and found it to provide a remarkably good fit with model-based estimates of the mean and variances in close alignment with the robust counterparts.

We have developed a framework for the design and analysis of large scale trials involving marked point processes where there is inter-individual variation in the rate of events. The events are the occasions in which a therapeutic intervention is warranted and the trial is planned such that the same intervention is applied each time an event occurs during the course of follow-up. We consider the setting where the response is available relatively soon after the therapy is administered and it does not affect the intensity for future events. This is suitable for the study of chronic conditions with recurrent episodic flares or acute exacerbations that warrant treatment and in the introduction we listed example settings.

We have assumed that there is no association between the number of donations and the risk of adverse events. In administrative data (not provided) we have found that adverse events can arise in any donation in a sequence of donations and hence they do not necessarily have an impact on the propensity to return for future donations. If this was a concern then one could consider the introduction of weights to the generalized estimating equations we provide to address this, or adopt a fully specified model for the multivariate binary response model and carry out likelihood analyses. Stopping rules of this sort are not problematic to likelihood analyses but the need to model the association between responses arises. The exposure status can also have an impact on the number of donations per donor so the cluster size can be informative. If our focus is on marginal population level effects, we can assume the working independence assumption even in the presence of informative cluster sizes ([McGee et al, 2020](#)) because GEE provides robust estimates of the marginal parameters even when the correlation structure within clusters is misspecified ([Fitzmaurice et al, 2012](#)).

Chapter 4

Two-Phase Designs for Biomarker Studies on Disease Progression

4.1 Introduction

Many modern cohort studies incorporate the creation of biobanks for the collection and storage of biospecimens such as serum or tissue samples ([Devereux et al, 2019](#)). Some biobanks are aimed at studying risk factors for a broad range of chronic diseases such as the U.K. Biobank ([Peakman and Elliott, 2008](#); [Collins, 2012](#)), the Canadian Longitudinal Study in Aging ([Raina et al, 2009](#)), and the Canadian Partnership for Tomorrow Project ([Dummer et al, 2018](#)). Other cohort studies may be designed to focus on particular conditions such as rheumatoid arthritis ([Dale et al, 2016](#)), diabetes ([Petersen et al, 2016](#)), or neurodegenerative diseases ([Reijs et al, 2015](#)). Biospecimens are usually taken upon recruitment to cohort studies and stored to be made available for future biomarker sub-studies which may aim to study the association between genetic, molecular or metabolomic features and disease onset or complications arising from disease ([Bebu et al, 2021](#)). In massive cohort studies it is practically and economically infeasible to assay biospecimens from all individuals and in some settings it is also important to preserve samples for future biomarker studies. In situations where the interest is in the association between an expensive biomarker and the disease outcome, a two-phase study design provides a cost-efficient way for estimating the biomarker effect. The first phase (phase I) of such a study involves the collection of inexpensive covariates (e.g. demographic and clinical variables) along with outcomes often observed through prospective followup of participants. The second phase (phase II) involves assaying biospecimens for a subset of selected individuals in order to

measure the expensive covariate of interest. The size of the phase II sub-sample (or its expected value) is set by budgetary constraints and the key design challenge is to choose the sub-sample according to a selection process which optimizes the precision of the biomarker effect. This design will be influenced by the method of analysis which will address the incomplete nature of the biomarker data through likelihood, conditional likelihood, inverse probability weighted estimating functions or other methods.

Since the phase I sample data can be used to guide the selection in phase II sample, these designs can feature response-dependent selection models. Many studies employ an extreme-outcome sampling design, where individuals with the highest or lowest values of a continuous outcome are selected; see [Chen and Li \(2011\)](#), [Li et al \(2011\)](#) and [Lin et al \(2013\)](#). [Breslow and Cain \(1988\)](#) developed the methodology for two-phase design for binary outcomes modeled via logistic regression. For failure time outcomes, [Ding et al \(2014\)](#) suggested selecting cases with the largest or smallest values of failure times and [Lawless \(2018\)](#) suggested selecting subjects with the smallest failure times or largest censored times. [Derkach et al \(2015\)](#) developed a sampling scheme in which extreme score-based residuals guide phase II selection. [Tao et al \(2020\)](#) extended this approach to describe a general framework for efficient two-phase design that can be applied to settings with continuous, discrete, or censored outcomes. [Zhou et al \(2020\)](#) discuss two-phase sampling schemes for failure time data when the failure time is subject to interval-censoring and [Mao and Cook \(2023\)](#) consider the setting with current status data.

Despite the extensive methodology for two-phase designs with response-dependent sub-sampling, relatively little work has been done for settings where the model for the disease process is more complex than a censored failure time. In many chronic conditions the disease involves progression through different stages and interest lies in examining the association between possible risk factors and progression rates more broadly. When the disease process involves a series of distinct stages it can be naturally modelled as a multi-state process ([Cook and Lawless, 2018](#); [Andersen and Ravn, 2023](#)). When the transition times between states are observed precisely then methods based on failure time processes can be readily adapted, but unique challenges arise when transition times between disease states are not observable. Examples where this is the case include studies of the progression of retinopathy in diabetes ([Ghamdi, 2020](#)), cirrhosis in liver disease ([Rosselli et al, 2013](#)), and dementia in neurological studies ([Albert et al, 2018](#)). In these settings imaging is required to assess the extent of damage and to infer whether transitions to more advanced disease states occurred; for other conditions the extent of damage may be determined by clinical examination. Likelihood construction for such data is given for Markov processes by [Kalbfleisch and Lawless \(1985\)](#) who also describe a Fisher-scoring algorithm for parameter estimation. This method is implemented in the package `msm` by [Jackson \(2011\)](#) in

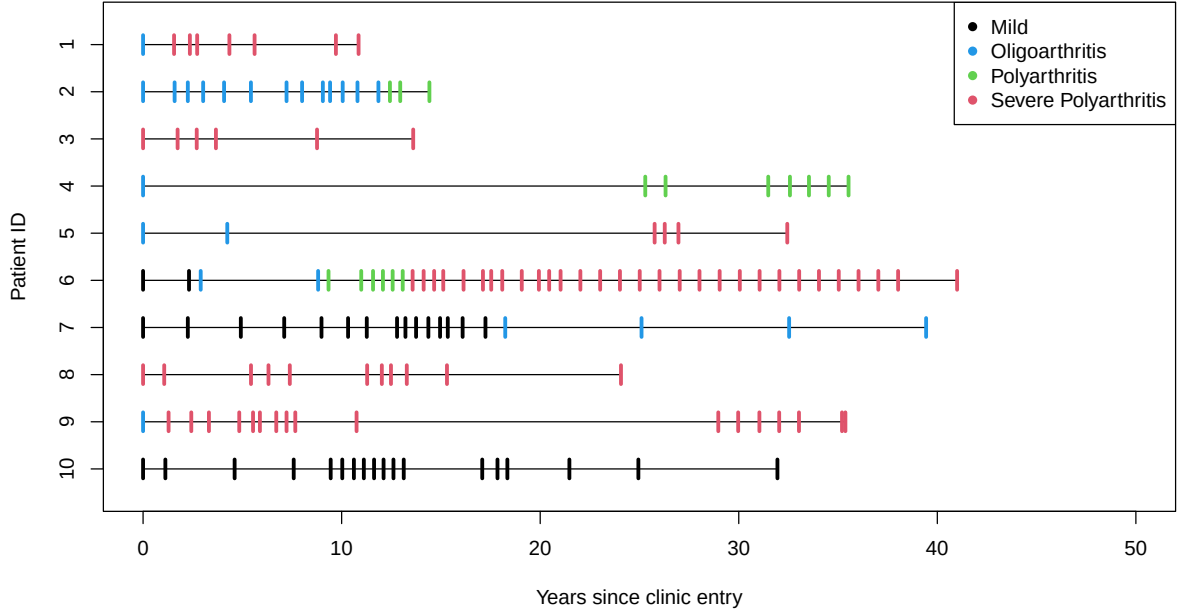


Figure 4.1: Ten sample timelines depicting disease progression obtained from follow-up visits since clinic entry on joint damage in patients with psoriatic arthritis (PsA) from the University of Toronto Psoriatic Arthritis Cohort.

R.

We consider the design of two-phase studies aimed at selecting biospecimens for assay in order to efficiently measure the biomarker effect on the transition intensities in progressive disease processes. The work is motivated by research at the University of Toronto Centre for Prognostic Studies in Rheumatic Disease which maintains a registry of patients with psoriatic arthritis (PsA) ([Gladman and Chandran, 2011](#)). Upon recruitment to the registry blood samples are taken and stored in a biobank, and baseline data are collected on demographic, clinical and lifestyle variables. Patients are then followed according to a protocol which involves annual clinical examination and biannual radiographic examination. Upon radiographic examination each of 64 joints is assessed for its degree of damage according to the modified Steinbrocker score ([Rahman et al, 1998](#)) and the number of damaged joints is recorded. Based on this number patients can be classified as having very mild disease (0 damaged joint), oligoarthritis (1-4 damaged joints) polyarthritis (5-9 damaged joints)

and severe polyarthritis (> 9 damaged joints). Figure 4.1 presents sample data for 10 individuals from the University of Toronto Psoriatic Arthritis Cohort. The length of each line indicates the duration each individual was followed since their initial clinic entry, and the vertical ticks represent visits when radiographic examinations were conducted to assess joint damage. Different colors in Figure 4.1 represent distinct states of joint damage. We consider a multistate process based on such a progressive model and aim to assess the role of the HLA marker B27 on the risk of progression across these states.

The remainder of this Chapter is structured as follows. In Section 4.2 we describe a continuous-time progressive disease process and define the notations for the setting where observations are intermittent. A common coefficient of interest across transition states for the covariates is considered. In Section 4.3 we present the likelihood formulation where an expensive covariate of key interest is missing. The analyses methods including maximum likelihood, conditional maximum likelihood, and inverse probability weighting are discussed. Section 4.4 introduces the approximately optimal score residual-based phase II sub-sampling scheme proposed by Tao et al (2020) along with the stratified phase II sub-sampling approach. In Section 4.5, we conduct a simulation study to assess the efficiency of the effect of the biomarker obtained from different sub-sampling frameworks using different analyses approaches. We demonstrate the efficiency gains through score residual-based sub-sampling strategy. In Section 4.6 we present a motivating example involving patients with psoriatic arthritis where we investigate the effect of HLA-B27 on the number of damaged joints in a two-phase design. Concluding remarks and directions for future research are given in Section 4.7.

4.2 Intermittent Observation of Progressive Processes

Consider a multistate process with state space $\mathcal{S} = \{0, 1, \dots, K\}$ where state 0 represents an initial healthy state and states $1, 2, \dots, K$ progressively more advanced disease states. Figure 4.2 is a schematic diagram representing the particulars of this disease process in which transitions can occur from state $k - 1$ to state k , $k = 1, \dots, K$. There are many chronic diseases which can be represented by such processes including liver cirrhosis in hepatitis, functional declines in musculoskeletal disease, or nephropathy in kidney disease. In other contexts with large K one may define the states as above based on the cumulative number of events such as fractures in osteoporosis, falls in gerontology, or internal bleeds in neurology. Let $Z(t)$ denote the disease state occupied at time $t \geq 0$, where $Z(t) \in \mathcal{S}$, and $\{Z(t), 0 \leq t\}$ represent the stochastic disease process. Let $\mathbf{X}_2$ denote a vector of inexpensive covariates which are easily available upon recruitment to a cohort and X_1

represent a biomarker of interest, with $\mathbf{X} = (X_1, \mathbf{X}_2')'$. The history for the disease process at t is denoted by $\mathcal{H}(t) = \{Z(s), 0 \leq s < t, \mathbf{X}\}$.

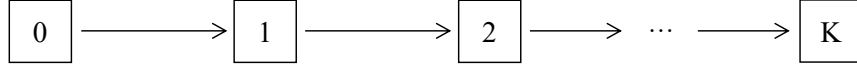


Figure 4.2: A progressive multistate model with $K + 1$ states.

The stochastic nature of the disease process is specified through the intensity functions (Cook and Lawless, 2018)

$$\lim_{\Delta t \downarrow 0} \frac{P(Z(t + \Delta t^-) = k + 1 \mid Z(t^-) = k, \mathcal{H}(t))}{\Delta t} = \lambda_k(t \mid \mathcal{H}(t)), \quad k = 0, 1, \dots, K - 1.$$

For modulated Markov models with multiplicative covariate effects (Cook and Lawless, 2018) the intensities have the form

$$\lambda_k(t \mid \mathcal{H}(t)) = \lambda_{k0}(t; \boldsymbol{\alpha}) \exp(\beta_1 X_1 + \boldsymbol{\beta}_2' \mathbf{X}_2), \quad (4.1)$$

where $\lambda_{k0}(t; \boldsymbol{\alpha})$ is the baseline intensity function indexed by $\boldsymbol{\alpha}$, $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2')'$ is a vector of regression coefficients, and $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$. Our primary focus is on β_1 which is the effect of the biomarker on the transition intensities for the disease processes.

The disease process is under incomplete observation due to intermittent observation over an administrative censoring period and possible random right censoring from loss to follow-up. Let $(0, \tau]$ denote the planned period of observation with administrative censoring time τ , C_0 be a random loss to follow-up time, and $C = \min(C_0, \tau)$ denote the net censoring time. We define $Y(t) = I(t \leq C)$ to indicate whether the individual is under observation at study time t . We let $A(t)$ count the number of clinic visits over $(0, t]$ with $dA(t) = 1$ if the visit occurred at time t and $dA(t) = 0$ otherwise. If $d\bar{A}(t) = Y(t)dA(t)$ then $\{\bar{A}(t), 0 < t\}$ is the counting process for visits at which the disease state is assessed where $\bar{A}(t) = \int_0^t d\bar{A}(s)$ is the observed number of assessments over $(0, t]$. In what follows we assume that the conditionally independent visit process (CIVP) assumption of Cook and Lawless (2021) is satisfied. Let $a_0 < a_1 < \dots < a_m$ be the realized visit times (i.e. $d\bar{A}(t) = 1$ if $t = a_r$ for some $r \in \{1, \dots, m\}$ but $d\bar{A}(t) = 0$ otherwise) so that total number of visits over $(0, \tau]$ is $m = \bar{A}(\tau) = \int_0^\tau d\bar{A}(s)$. Complete data on the disease process and covariates for this hypothetical individual can be represented as $\mathcal{D} = \{(Z(a_r), a_r), r = 0, 1, \dots, m; \mathbf{X}\}$. With this complete data and under the CIVP assumption (Cook and Lawless, 2021) the

likelihood conditional on the covariate $\mathbf{X}$ is

$$P(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) \propto \prod_{r=1}^m P(Z(a_r) \mid a_r, Z(a_{r-1}), \mathbf{X}; \boldsymbol{\theta}) .$$

In the framework of two-phase designs, the covariate X_1 is not available in the phase I sample so the data are $\mathcal{D}_I = \{(Z(a_r), a_r), r = 0, 1, \dots, m; \mathbf{X}_2\}$. Individuals are to be chosen to form a phase II sub-sample in which X_1 is measured by assaying the stored biospecimens. We let R indicate the selection such that $R = 1$ if the individual has their biospecimens assayed and X_1 is measured (i.e. they are included in the phase II sub-sample) and $R = 0$ otherwise. The selection of phase II subsamples may depend on the full phase I data or only the observed data $\mathcal{D}_I$ for the respective individual, and in either case X_1 is missing at random (MAR) (Little and Rubin, 2019). Likelihood estimates retain good properties when covariates are MAR but weighting may be needed for other approaches as we discuss next. We propose and evaluate phase II selection methods and models in terms of the precision of the estimator characterizing the biomarker effect on disease progression ($\hat{\beta}_1$) when the disease process is under intermittent observation. The various methods of analysis and sub-sampling are compared in settings where we respect budgetary constraints. In Section 4.3 we first discuss methods of estimation for multistate processes under intermittent observation when covariates are incomplete.

4.3 Analysis Methods for Two-Phase Studies

4.3.1 Maximum Likelihood

Maximum likelihood estimation is the most efficient method for dealing with incomplete covariates but it does require specification of the covariate distribution which can be challenging when X_1 is continuous or $\mathbf{X}_2$ is high dimensional and contains a combination of categorical and continuous covariates. While in general this can raise robustness concerns in what follows we consider discrete covariates. Let $P(\mathbf{X}; \boldsymbol{\gamma})$ denote the joint distribution of the covariates which can be factored as

$$P(\mathbf{X}; \boldsymbol{\gamma}) = P(X_1 \mid \mathbf{X}_2; \boldsymbol{\gamma}_1) P(\mathbf{X}_2; \boldsymbol{\gamma}_2) ,$$

with $\boldsymbol{\gamma} = (\boldsymbol{\gamma}'_1, \boldsymbol{\gamma}'_2)'$. The joint probability of inexpensive covariates $P(\mathbf{X}_2; \boldsymbol{\gamma}_2)$ can be fitted based on the phase I sample and if covariates are discrete this is straightforward. The full set of parameters is denoted by $\boldsymbol{\vartheta} = (\boldsymbol{\theta}', \boldsymbol{\gamma}')'$.

Consider a sample of observations from n independent processes yielding phase I data $\{(Z(a_{ir}), a_{ir}), r = 0, \dots, m_i, \mathbf{X}_{i2}\}$ for process i , $i = 1, \dots, n$. Let $R_i = 1$ if individual i is selected for inclusion in the phase II sample in which case we observe $\mathcal{D}_i^\circ = \{(Z(a_{ir}), a_{ir}), r = 0, \dots, m_i, \mathbf{X}_i\}$ if $R_i = 1$ and $\mathcal{D}_i^\circ = \{(Z(a_{ir}), a_{ir}), r = 0, \dots, m_i, \mathbf{X}_{i2}\}$ if $R_i = 0$, $i = 1, \dots, n$. This gives the data after phase II selection as $\{\mathcal{D}_i^\circ, i = 1, \dots, n\}$ yielding the observed data likelihood

$$L(\boldsymbol{\vartheta}) = \prod_{i=1}^n \{P(\mathcal{D}_i | \mathbf{X}_i; \boldsymbol{\theta}) P(X_{i1} | \mathbf{X}_{i2}; \boldsymbol{\gamma}_1) P(\mathbf{X}_{i2}; \boldsymbol{\gamma}_2)\}^{R_i} \times \left\{ \int P(\mathcal{D}_i | \mathbf{X}_i; \boldsymbol{\theta}) P(X_{i1} | \mathbf{X}_{i2}; \boldsymbol{\gamma}_1) P(\mathbf{X}_{i2}; \boldsymbol{\gamma}_2) dx_1 \right\}^{1-R_i},$$

where

$$P(\mathcal{D}_i | \mathbf{X}_i; \boldsymbol{\theta}) = \prod_{r=1}^{m_i} \prod_{k,l=1}^K P(Z(a_{ir}) = l | Z(a_{i,r-1}) = k, \mathbf{X}_i; \boldsymbol{\theta})^{\delta_{iklr}} = \prod_{r=1}^{m_i} \prod_{k,l=1}^K P_{iklr}(\mathbf{X}; \boldsymbol{\theta})^{\delta_{iklr}},$$

and $\delta_{iklr} = \mathbb{I}(Z(a_{i,r-1}) = k, Z(a_{ir}) = l)$ denotes the indicator of subject i in state k at $a_{i,r-1}$ and l at a_{ir} . The transition probability $P_{iklr}(\mathbf{X}; \boldsymbol{\theta}) = P(Z(a_{ir}) = l | Z(a_{i,r-1}) = k, \mathbf{X}_i; \boldsymbol{\theta})$ can be obtained from the $(k+1, l+1)$ entry of the transition probability matrix $P_r(\mathbf{X}; \boldsymbol{\theta})$ (Cox, 2017) where

$$P_r(\mathbf{X}; \boldsymbol{\theta}) = \exp \{Q(\mathbf{X}; \boldsymbol{\theta})(a_r - a_{r-1})\}, \quad (4.2)$$

and $Q(\mathbf{X}; \boldsymbol{\theta})$ is the transition intensity matrix (Cox, 2017) with $(k+1, l+1)$ entry being $\lambda_{kl}(t | \mathcal{H}(t^-))$, $k, l = 0, \dots, K$, with

$$\lambda_{kl}(t | \mathcal{H}(t^-)) = \begin{cases} \lambda_k(t | \mathcal{H}(t^-)), & k = 0, 1, \dots, K-1, l = k+1 \\ 0, & l > k+1, l < k \\ -\sum_{j \neq k} \lambda_{kj}(t | \mathcal{H}(t^-)), & l = k. \end{cases}$$

Hence the full log-likelihood is written as

$$\begin{aligned} \ell(\boldsymbol{\vartheta}) &= \sum_{i=1}^n \ell_i(\boldsymbol{\vartheta}) = \sum_{i=1}^n R_i \{ \log P(\mathcal{D}_i | \mathbf{X}_i; \boldsymbol{\theta}) + \log P(X_{i1} | \mathbf{X}_{i2}; \boldsymbol{\gamma}_1) \} \\ &\quad + (1 - R_i) \left\{ \log \int P(\mathcal{D}_i | \mathbf{X}_i; \boldsymbol{\theta}) P(X_{i1} | \mathbf{X}_{i2}; \boldsymbol{\gamma}_1) dx_1 \right\} + \log P(\mathbf{X}_{i2}; \boldsymbol{\gamma}_2). \end{aligned} \quad (4.3)$$

The estimate $\hat{\boldsymbol{\vartheta}}$ is obtained by solving the score equations

$$\mathbf{S}(\boldsymbol{\vartheta}) = \sum_{i=1}^n \mathbf{S}_i(\boldsymbol{\vartheta}) = \left(\sum_{i=1}^n \mathbf{S}'_{i1}(\boldsymbol{\vartheta}), \sum_{i=1}^n \mathbf{S}'_{i2}(\boldsymbol{\vartheta}), \sum_{i=1}^n \mathbf{S}'_{i3}(\boldsymbol{\gamma}_2) \right)' = \mathbf{0},$$

where

$$\mathbf{S}_{i1}(\boldsymbol{\vartheta}) = \frac{\partial \ell_i(\boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}} = R_i \mathcal{S}_{i1}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}) + (1 - R_i) \mathbb{E}_{X_{i1} \mid \mathcal{D}, \mathbf{X}_2} \{ \mathcal{S}_{i1}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}); \boldsymbol{\vartheta}_1 \}, \quad (4.4)$$

$$\mathbf{S}_{i2}(\boldsymbol{\vartheta}) = \frac{\partial \ell_i(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_1} = R_i \mathcal{S}_{i2}(X_{i1} \mid \mathbf{X}_{i2}; \boldsymbol{\gamma}_1) + (1 - R_i) \mathbb{E}_{X_{i1} \mid \mathcal{D}, \mathbf{X}_2} \{ \mathcal{S}_{i2}(X_{i1} \mid \mathbf{X}_{i2}; \boldsymbol{\gamma}_1); \boldsymbol{\vartheta}_1 \}, \quad (4.5)$$

and

$$\mathbf{S}_{i3}(\boldsymbol{\gamma}_2) = \frac{\partial \ell_i(\boldsymbol{\vartheta})}{\partial \boldsymbol{\gamma}_2} = \mathcal{S}_{i3}(\mathbf{X}_{i2}; \boldsymbol{\gamma}_2). \quad (4.6)$$

The complete score $\mathcal{S}_{i1}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}) = \partial \log P(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ can be computed using the matrix decomposition given $\boldsymbol{\theta}$ as discussed by [Kalbfleisch and Lawless \(1985\)](#); $\mathcal{S}_{i2}(X_{i1} \mid \mathbf{X}_{i2}; \boldsymbol{\gamma}_1) = \partial \log P(X_{i1} \mid \mathbf{X}_{i2}; \boldsymbol{\gamma}_1) / \partial \boldsymbol{\gamma}_1$ and $\mathcal{S}_{i3}(\mathbf{X}_{i2}; \boldsymbol{\gamma}_2) = \partial \log P(\mathbf{X}_{i2}; \boldsymbol{\gamma}_2) / \partial \boldsymbol{\gamma}_2$ usually have a simple form given the parameters. Details regarding the derivation of (4.4) are provided in [Appendix C.1](#).

From (4.4) and (4.5) we see that the observed data score for individuals not selected into phase II (i.e. $R_i = 0$) can be expressed as the conditional expectation of the complete data score ([Louis, 1982](#)). Consequently once the phase II sample has been selected, the expectation-maximization (EM) algorithm ([Dempster et al, 1977](#)) can be used to estimate $\boldsymbol{\vartheta}$. Let

$$\eta_{i1}^{(r)} = P(X_{i1} = 1 \mid \mathcal{D}_i, \mathbf{X}_{i2}; \boldsymbol{\vartheta}^{(r-1)})$$

be the estimated probability from the r th iteration. The parameters can be estimated by solving equation $U^{(r)}(\boldsymbol{\vartheta}) = \mathbf{0}$ iteratively until the convergence of $\hat{\boldsymbol{\vartheta}}$, where $U^{(r)}(\boldsymbol{\vartheta})$ is

$$\begin{pmatrix} \sum_{i=1}^n R_i \mathcal{S}_{i1}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}) + (1 - R_i) \left\{ \mathcal{S}_{i1}(\mathcal{D}_i \mid X_{i1} = 1, \mathbf{X}_{i2}; \boldsymbol{\theta}) \eta_{i1}^{(r)} + \mathcal{S}_{i1}(\mathcal{D}_i \mid X_{i1} = 0, \mathbf{X}_{i2}; \boldsymbol{\theta}) (1 - \eta_{i1}^{(r)}) \right\} \\ \sum_{i=1}^n R_i \mathcal{S}_{i2}(X_{i1} \mid \mathbf{X}_{i2}; \boldsymbol{\gamma}_1) + (1 - R_i) \left\{ \mathcal{S}_{i2}(X_{i1} = 1 \mid \mathbf{X}_{i2}; \boldsymbol{\gamma}_1) \eta_{i1}^{(r)} + \mathcal{S}_{i2}(X_{i1} = 0 \mid \mathbf{X}_{i2}; \boldsymbol{\gamma}_1) (1 - \eta_{i1}^{(r)}) \right\} \end{pmatrix}.$$

The EM algorithm is particularly flexible and robust when dealing with complex models but the convergence of the estimates can be slow due to iterative computation of expectations and maximizations. Alternatively $\hat{\boldsymbol{\vartheta}}$ can be obtained through direct maximization of the log-likelihood function in (4.3). This approach may be more computationally efficient than

the EM algorithm if the likelihood function is not overly complicated to evaluate and optimize. Under the standard regularity condition the $\hat{\boldsymbol{\vartheta}}$ is asymptotically normal with

$$\sqrt{n} \left(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta} \right) \xrightarrow{d} \text{MVN} \left(0, \mathbb{E} \{ \mathbf{S}_i(\boldsymbol{\vartheta})' \mathbf{S}_i(\boldsymbol{\vartheta}) \}^{-1} \right) \quad \text{as } n \rightarrow \infty. \quad (4.7)$$

In practice the asymptotic variance can be estimated using the empirical averages, i.e.,

$$\widehat{\text{cov}}(\sqrt{n}(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})) = \left\{ \frac{1}{n} \sum_{i=1}^n \mathbf{S}_i(\boldsymbol{\vartheta})' \mathbf{S}_i(\boldsymbol{\vartheta}) \right\}^{-1} \Big|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\vartheta}}}.$$

4.3.2 Conditional Maximum Likelihood

The maximum likelihood approach involves estimating both the parameters of interest and nuisance parameters. An alternative likelihood-based method is the conditional maximum likelihood approach, which does not require the modeling of the nuisance parameters (Berger et al, 1999). We assume that the initial state is 0 for all subjects, the conditional maximum likelihood can be written as

$$L_C(\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^n P(\mathcal{D}_i \mid \mathbf{X}_i, R_i = 1; \boldsymbol{\theta}, \boldsymbol{\psi})^{R_i} = \prod_{i=1}^n \left\{ \frac{P(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}) \pi_i}{\sum_z P(Z(a_{im_i}) = z \mid Z(0) = 0, \mathbf{X}_i; \boldsymbol{\theta}) \pi_i} \right\}^{R_i}$$

where $\pi_i = P(R_i = 1 \mid Z(a_{im_i}), \mathbf{X}_{i2}; \boldsymbol{\psi})$ and $\boldsymbol{\psi}$ is the vector of parameters governing the sampling probability. Note that the nuisance distribution $P(X_{i1} \mid \mathbf{X}_{i2}; \boldsymbol{\gamma}_1)$ is eliminated from the conditional maximum likelihood but the sampling probability needs to be estimated if it is not known and it can take the form of the disease process and inexpensive covariates which are available for all subjects, e.g.,

$$\text{logit } \pi_i = \psi_0 + \psi_1 Z(a_{im_i}) + \psi_2' \mathbf{X}_{i2} + \psi_3' Z(a_{im_i}) \mathbf{X}_{i2}. \quad (4.8)$$

The parameters $\hat{\boldsymbol{\psi}}$ can be obtained by fitting a logistic regression. The sampling weights are then estimated by $\hat{\pi}_i = \hat{\pi}_i(\hat{\boldsymbol{\psi}})$ or through the re-estimation by solving equation (4.9) which uses the information of all individuals in the phase I sample (Yang et al, 2022),

$$\sum_{i=1}^n U_{i0}(\boldsymbol{\psi}) = \sum_{i=1}^n \frac{\partial \pi_i}{\partial \boldsymbol{\psi}} \frac{R_i - \pi_i}{\pi_i(1 - \pi_i)} = 0. \quad (4.9)$$

The estimates $\hat{\boldsymbol{\theta}}$ can be obtained by solving the score equations from the phase II

subsamples,

$$\sum_{i=1}^n S_{iC}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\psi}) = \sum_{i=1}^n R_i \frac{\partial}{\partial \boldsymbol{\theta}} \log P(\mathcal{D}_i \mid \mathbf{X}_i, R_i = 1; \boldsymbol{\theta}, \boldsymbol{\psi}) = 0. \quad (4.10)$$

Substituting $\hat{\pi}_i$ for π_i in (4.10) and with the regularity condition (Lawless et al, 1999; Andersen, 1970) $\hat{\boldsymbol{\theta}}$ is asymptotically normal with

$$\sqrt{n} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right) \xrightarrow{d} \text{MVN} \left(0, \mathcal{A}_C^{-1} (\mathcal{B}_C - \Gamma_C \text{H} \Gamma_C') (\mathcal{A}_C^{-1})' \right) \quad \text{as } n \rightarrow \infty,$$

where $\mathcal{A}_C = \mathbb{E}\{-\partial S_{iC}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\psi}) / \partial \boldsymbol{\theta}\}$, $\mathcal{B}_C = \mathbb{E}\{S_{iC}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\psi}) S_{iC}'(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\psi})\}$, $\Gamma_C = \mathbb{E}\{-\partial S_{iC}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\psi}) / \partial \boldsymbol{\psi}\}$, and $\text{H} = \mathbb{E}\{U_{i0}(\boldsymbol{\psi}) U_{i0}'(\boldsymbol{\psi})\}$.

4.3.3 Inverse Probability Weighted Estimating Equations

The biomarker X_1 is only available for phase II subsamples which introduces a missing data problem. Inverse probability weighting (IPW) is commonly used to handle missing data by assigning each individual a weight based on the inverse of their sampling probability (i.e. $1/\pi_i$). This approach allows for valid inferences by accounting for the unequal probabilities of selection. The estimate $\hat{\boldsymbol{\theta}}$ is then obtained by solving the inverse probability weighted estimating equations

$$\sum_{i=1}^n U_{i1}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\psi}) = \sum_{i=1}^n \frac{R_i}{\pi_i} S_{i1}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}) = 0, \quad (4.11)$$

where

$$S_{i1}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log P(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}).$$

We again replace π_i by $\hat{\pi}_i$ in (4.11) and under the regularity condition (Lawless et al, 1999; Robins et al, 1995) $\hat{\boldsymbol{\theta}}$ is asymptotically normal with

$$\sqrt{n} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right) \xrightarrow{d} \text{MVN} \left(0, \mathcal{A}^{-1} (\mathcal{B} - \Gamma \text{H} \Gamma') (\mathcal{A}^{-1})' \right) \quad \text{as } n \rightarrow \infty,$$

where $\mathcal{A} = \mathbb{E}\{-\partial U_{i1}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\psi}) / \partial \boldsymbol{\theta}\}$, $\mathcal{B} = \mathbb{E}\{U_{i1}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\psi}) U_{i1}'(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\psi})\}$, and $\Gamma = \mathbb{E}\{-\partial U_{i1}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\psi}) / \partial \boldsymbol{\psi}\}$. Note that if the sampling probability is known rather than estimated, the asymptotic variance would be $\mathcal{A}^{-1} \mathcal{B} (\mathcal{A}^{-1})'$ (McIsaac and Cook, 2017) where $\boldsymbol{\psi}$ is treated as the true parameter and can be omitted in the score function,

i.e. $U_{i1}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\psi}) = U_{i1}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta})$.

The IPW approach is straightforward to implement as it also does not require modelling the nuisance distribution $P(X_{i1} \mid \mathbf{X}_{i2}; \boldsymbol{\gamma}_1)$. However its performance highly depends on the correct specification of the sampling probability model π_i . If π_i is misspecified, the IPW method may generate extreme weights for certain individuals, potentially leading to biased estimates with high variance. Additionally IPW tends to be less efficient compared to the maximum likelihood method because it only utilizes partial information from phase I data ([Chen and Lumley, 2022](#)).

4.4 Two-Phase Sub-Sampling Strategies

4.4.1 Residual Dependent Phase II Subsampling

Here we consider the optimal two-phase design based on the maximum likelihood estimation. We let the asymptotic variance of $\hat{\beta}_1$ from (4.7) be V_{β_1} , i.e. $V_{\beta_1} = \text{var}(\sqrt{n}(\hat{\beta}_1 - \beta_1))$. The aim is to find the optimal two-phase design which minimizes V_{β_1} subject to the budget constraint

$$P(R_i = 1) = p_c, \quad (4.12)$$

where p_c is the fixed phase II sampling fraction according to the study budget, $0 < p_c \leq 1$. The phase II sample size can then be determined as $n_2 = np_c$.

We assume a time-homogeneous Markov model for the state transition so the transition intensity from state k to $k + 1$ in (4.1) can be simplified to

$$\lambda_k(\mathbf{X}; \boldsymbol{\theta}) = \lambda_{k0} e^{X_1 \beta_1 + \mathbf{X}_2' \boldsymbol{\beta}_2},$$

where $\boldsymbol{\theta} = (\lambda_{k0}, \beta_1, \boldsymbol{\beta}_2')'$, $k = 0, 1, \dots, K - 1$.

The transition probability $P_r(\mathbf{X}; \boldsymbol{\theta})$ in (4.2) then has a close form expression with the (k, l) entry given by ([Satten, 1999](#))

$$P_{klr}(\mathbf{X}) = P(Z(a_r) = l \mid Z(a_{r-1}) = k, \mathbf{X}; \boldsymbol{\theta}) = \begin{cases} \sum_{j=k}^l C_{kjl} e^{-\lambda_j(\mathbf{X}; \boldsymbol{\theta})(a_r - a_{r-1})}, & k \leq l \\ 0, & k > l, \end{cases}$$

where

$$C_{kjl} = \frac{\prod_{s=k}^{l-1} \lambda_s(\mathbf{X}; \boldsymbol{\theta})}{\prod_{s=k, s \neq j}^l (\lambda_s(\mathbf{X}; \boldsymbol{\theta}) - \lambda_j(\mathbf{X}; \boldsymbol{\theta}))} = \frac{\prod_{s=k}^{l-1} \lambda_{s0}}{\prod_{s=k, s \neq j}^l (\lambda_{s0} - \lambda_{j0})}, \quad k \leq j \leq l,$$

and $C_{jjj} = 1$.

The calculations of V_{β_1} can be greatly simplified and its expression is analytically tractable under the assumption that $\beta_1 = o(1)$ (Tao et al, 2020). Tao et al (2020) argue that this assumption is reasonable as biomarkers often exhibit a small effect on outcomes in many studies. It is essential to assess the efficiency of the biomarker estimator particularly when the effect size is indeed small. Moreover many empirical findings show that the two-phase design with the residual-dependent sampling under $\beta_1 = o(1)$ assumption yields estimators with higher efficiency, even when the true β_1 is not small (Tao et al, 2020; Mao and Cook, 2023).

Under the assumption $\beta_1 = o(1)$, we can write V_{β_1} (Tao et al, 2020) as

$$V_{\beta_1} = [\Sigma_1 + \mathbb{E}\{R \cdot \text{var}(M_\mu \mid R = 1, \mathbf{X}_2) \text{var}(X_1 \mid \mathbf{X}_2)\}]^{-1}, \quad (4.13)$$

where Σ_1 is the fisher information of β_1 which does not depend on R , and M_μ is the derivative of $\log P(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta})$ with respect to $\mu = X_1\beta_1 + \mathbf{X}_2'\boldsymbol{\beta}_2$. Furthermore M_μ can be written as

$$\begin{aligned} M_\mu &= \frac{\partial \log P(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta})}{\partial \mu} \\ &= \sum_{r=1}^m \sum_{k=1}^K \sum_{l=1}^K -\delta_{klr}(a_r - a_{r-1}) \frac{\sum_{j=k}^l \lambda_j(\mathbf{X}; \boldsymbol{\theta}) P_{klr}^{(j)}(\mathbf{X})}{P_{klr}(\mathbf{X})}, \quad k \leq l, \end{aligned} \quad (4.14)$$

where $P_{klr}^{(j)}(\mathbf{X}) = C_{kjl} \exp\{-\lambda_j(\mathbf{X}; \boldsymbol{\theta})(a_r - a_{r-1})\}$. It can be shown that $\mathbb{E}(M_\mu) = 0$ and M_μ can be viewed as the residuals of the complete data score (Mao and Cook, 2023). The details of related derivations are provided in Appendix C.2.

Since Σ_1 in (4.13) does not depend on the phase II sampling, minimizing (4.13) is equivalent to the maximization of

$$\mathbb{E}\{R \cdot \text{var}(M_\mu \mid R = 1, \mathbf{X}_2) \text{var}(X_1 \mid \mathbf{X}_2)\} \quad (4.15)$$

subject to the constraint in (4.12). The optimal or approximately optimal design (when β_1 is far from zero) refers to the sampling rule that selects individuals with the extreme values of $M_\mu \text{var}(X_1 \mid \mathbf{X}_2)^{1/2}$ in each stratum defined by $\mathbf{X}_2$ while satisfying the constraint in (4.12) (Tao et al, 2020). The value of $\text{var}(X_1 \mid \mathbf{X}_2)$ is typically unknown without historical data. Mao and Cook (2023) propose using an adaptive design where a preliminary sample is selected from phase I and X_1 is measured for this small sample to estimate $\text{var}(X_1 \mid \mathbf{X}_2)$. Alternatively $\text{var}(X_1 \mid \mathbf{X}_2)$ can be treated as constant assuming $X_1 \perp \mathbf{X}_2$.

This assumption is often reasonable particularly when X_1 is the generic biomarker and $\mathbf{X}_2$ represents the demographic characteristics that are available for all individuals (Mao and Cook, 2023). We refer this sampling rule as the score residual-dependent sampling as M_μ can be considered as a score-type residual derived from the complete data.

4.4.2 Stratified Phase II Subsampling

Balanced sampling is a stratified sampling technique that ensures the selected subjects are representative of the population in terms of key auxiliary variables. It enhances the efficiency of statistical estimators by selecting an equal number of subjects from each stratum. Let $C(Z(a_{im_i}), \mathbf{X}_{i2})$ denote the stratum for subject i where $C(\cdot)$ maps the realization of $(Z(a_{im_i}), \mathbf{X}'_{i2})$ to strata labeled $1, 2, \dots, n_s$. The mapping $C(Z(a_{im_i}), \mathbf{X}_{i2})$ determines the covariates that define the strata for subsampling. Stratified subsampling can then be implemented with the sampling probability $\pi_i = P(R_i = 1 \mid Z(a_{im_i}), \mathbf{X}_{i2}; \boldsymbol{\psi})$ as modeled in (4.8).

In some cases certain strata may have very small sizes which are insufficient for selecting the desired number of subsamples. For strata with the insufficient number of subjects we can either randomly select samples with replacement to obtain the required sample size, or select all available subjects from that stratum and compensate by selecting additional subjects from larger strata to reach the target sample size. More generally suppose we plan to select n_2 subjects from n_s strata for phase II measurement. In balanced sampling we aim to select n_2/n_s subjects equally from each stratum. However, when some strata sizes are smaller than n_2/n_s we can partition n_2 into n_s groups considering all possible combinations of sample sizes. These combinations are then ranked from the lowest variance (representing equal sample sizes across strata, which is ideal) to the highest variance. We next select the first combination where the sampling sizes are less than or equal to the available strata sizes. For example if we need to select 6 subjects using balanced sampling from two strata with sizes of 10 and 2 respectively. The possible partitions ranked from lowest to highest variance are (3, 3), (4, 2), (2, 4), (5, 1), (1, 5), (6, 0), (0, 6). In this case we would select 4 subjects from the larger stratum and 2 subjects from the smaller one. This ensures we meet the target sample size while maintaining the lowest variance in sample sizes among strata, thus achieving approximately balanced sampling.

4.5 Empirical Studies of Two-Phase Designs

Here we examine the empirical efficiency of the estimator of biomarker effect under different two-phase designs and estimation methods. We consider simple random sampling (SRS), balanced sampling (BAL), and score residual-dependent sampling (RES) with the biomarker effect estimated using the maximum likelihood approach (ML). Additionally we evaluate the simple random sampling and balanced sampling strategies with biomarker effect estimated using the conditional maximum likelihood (CML), and inverse probability weighting approaches (IPW). We consider a phase I sample of $n = 2,000$ subjects, from which we plan to select a subsample of size $n_2 = 300$ or 600 for phase II measurements of the expensive biomarker X_1 . Suppose the inexpensive covariate is binary, generated as $X_2 \sim \text{Bernoulli}(0.5)$, and the expensive biomarker is generated as $X_1 \sim \text{Bernoulli}(P(X_1 = 1 | X_2))$ where $P(X_1 = 1 | X_2) = \text{expit}(\gamma_0 + \gamma_1 X_2)$, $\gamma_1 = \log(2)$, and γ_0 can be solved to ensure $P(X_1 = 1) = 0.2$ or 0.5 for low and moderate probability of having positive biomarker respectively.

We consider the progressive disease process with three states, states 0, 1, and final state 2. Suppose the baseline transition intensities are λ_{00} (from state 0 to state 1) and λ_{10} (from state 1 to state 2), with $\lambda_{10} = 1.1\lambda_{00}$, reflecting a higher intensity from state 1 to final state 2. We set the probability of reaching the final state 2 by time $\tau = 1$ to $P(Z(\tau) = 2) = 0.4$ or 0.8 , corresponding to moderate and high risk of getting into the final state by the end of administrative time τ . The baseline transition intensities λ_{00} and λ_{10} can then be obtained by solving $P(Z(\tau) = 2) = 0.4$ or 0.8 , and the transition intensity is modeled as $\lambda_k(\mathbf{X}; \boldsymbol{\theta}) = \lambda_{k0} e^{\beta_1 X_1 + \beta_2 X_2}$, $k = 0, 1$, where we let $\beta_1 = 0, \log(1.25), \log(2)$, and $\beta_2 = \log(2)$. With this setup the transition intensity matrix is determined, and individual disease progression can be simulated using `sim.msm` function from the `msm` R package (Jackson, 2011). Furthermore we assume all subjects start in state 0 at the origin, and their clinic visit times follow an exponential distribution with rate $\lambda_a = 4$ resulting in an average visit interval of $1/4$ units of time. These clinic visit times (i.e. a_{ir}) can then be generated and corresponding disease states occupied at each clinic visit for each subject (i.e. $Z(a_{ir})$) can be determined, $r = 0, 1, \dots, m_i$.

With the simulated data we implement three sampling strategies in phase II. For the simple random sampling we randomly select n_2 subjects from phase I; for score residual-dependent sampling we select $n_2/4$ subjects with the highest and $n_2/4$ subjects with the lowest values of $M_\mu\{\text{var}(X_1 | \mathbf{X}_2)\}^{1/2}$ within each stratum defined by $X_2 = j$, where $\text{var}(X_1 | \mathbf{X}_2 = j) = P(X_1 = 1 | X_2 = j)(1 - P(X_1 = 1 | X_2 = j))$, $j = 0, 1$. For the balanced sampling we define 6 strata based on the state observed at the last visit (i.e., state 0, 1, or 2) and the binary covariate X_2 ; see Table 4.1. We select an equal number of sub-

jects randomly from each stratum. If strata sizes are insufficient, approximately balanced sampling is performed. The data generation process is repeated for $n_{sim} = 1,000$ iterations for each parameter configuration with different designs. The selection probabilities in the stratified subsampling designs and the estimation methods are modeled as

$$\text{logit } \pi_i = \psi_0 + \psi_1 Z(a_{m_i}) + \psi_2 X_{i2} + \psi_3 Z(a_{m_i}) X_{i2} .$$

Table 4.1: Tabulation of strata in empirical studies of two-phase design.

	stratum 1	stratum 2	stratum 3	stratum 4	stratum 5	stratum 6
$Z(\tau)$	0	0	1	1	2	2
X_2	0	1	0	1	0	1

Three analysis methods considered here yield consistent estimates, and the empirical and analytical standard errors of the estimated biomarker effect for each setting are summarized in Table 4.2. The empirical standard errors and analytical standard errors shown in brackets are closely aligned. Estimates obtained using the maximum likelihood method have higher efficiency compared to those from the other two methods. Score residual-dependent sampling selects more subjects from the strata 1, 2, 5, and 6, indicating a preference for subjects who are either in the initial state or the final state at the last visit before censoring; see Table C.1 in Appendix C.2. Among all frameworks score residual-dependent sampling with the maximum likelihood method achieves the highest efficiency of the estimates, confirming that it is an approximately optimal design. Balanced sampling can be less efficient than simple random sampling under some parameter settings as it selects an equal number of subjects from each stratum which leads to the over-selection of subjects in the middle state 1 (strata 3 and 4) at the final visit compared to simple random sampling and score residual-dependent sampling. The conditional maximum likelihood method with simple random sampling or balanced sampling approaches the efficiency of the maximum likelihood method as the phase II sample size increases, while the inverse probability weighting approach is the least efficient overall.

4.6 Joint Damage in Psoriatic Arthritis

Psoriatic arthritis (PsA) is a chronic immunological disease where the immune system mistakenly attacks healthy joints, causing joint inflammation and potentially leading to joint damage. The University of Toronto Psoriatic Arthritis Clinic was established in

Table 4.2: Empirical standard errors and analytical standard errors[†] of $\hat{\beta}_1$ obtained from maximum likelihood (ML), conditional maximum likelihood (CML), and inverse probability weighting (IPW) approaches for two-phase designs with simple random sampling (SRS), balanced sampling (BAL), and score residual-based sampling (RES); phase I sample size $n = 2000$, $n_{sim} = 1000$.

n_2	$P(Z(\tau) = 2)$	$P(X_1 = 1)$	β_1	ML			CML		IPW	
				SRS	BAL	RES	SRS	BAL	SRS	BAL
300	0.4	0.2	0	0.155 (0.153)	0.148 (0.149)	0.098 (0.097)	0.156 (0.154)	0.151 (0.150)	0.158 (0.153)	0.154 (0.153)
			0.22	0.148 (0.144)	0.145 (0.141)	0.096 (0.094)	0.148 (0.147)	0.148 (0.144)	0.149 (0.147)	0.151 (0.146)
			0.69	0.119 (0.120)	0.120 (0.118)	0.093 (0.093)	0.136 (0.139)	0.138 (0.136)	0.137 (0.138)	0.141 (0.138)
		0.5	0	0.129 (0.125)	0.128 (0.120)	0.081 (0.080)	0.129 (0.126)	0.130 (0.121)	0.129 (0.126)	0.132 (0.125)
			0.22	0.126 (0.124)	0.116 (0.120)	0.078 (0.080)	0.129 (0.127)	0.119 (0.122)	0.130 (0.127)	0.122 (0.126)
			0.69	0.117 (0.118)	0.114 (0.115)	0.084 (0.085)	0.129 (0.133)	0.126 (0.128)	0.130 (0.133)	0.130 (0.133)
	0.8	0.2	0	0.121 (0.123)	0.107 (0.107)	0.069 (0.070)	0.122 (0.126)	0.123 (0.119)	0.124 (0.123)	0.136 (0.131)
			0.22	0.114 (0.117)	0.105 (0.103)	0.068 (0.069)	0.119 (0.124)	0.129 (0.114)	0.121 (0.121)	0.137 (0.133)
			0.69	0.095 (0.097)	0.090 (0.091)	0.094 (0.095)	0.120 (0.127)	0.149 (0.110)	0.120 (0.120)	0.147 (0.142)
		0.5	0	0.098 (0.099)	0.086 (0.088)	0.056 (0.057)	0.099 (0.101)	0.097 (0.097)	0.101 (0.100)	0.102 (0.106)
			0.22	0.095 (0.093)	0.084 (0.084)	0.056 (0.057)	0.100 (0.101)	0.099 (0.097)	0.102 (0.100)	0.109 (0.106)
			0.69	0.070 (0.070)	0.070 (0.068)	0.058 (0.056)	0.099 (0.102)	0.109 (0.098)	0.099 (0.101)	0.112 (0.111)
600	0.4	0.2	0	0.105 (0.107)	0.108 (0.105)	0.071 (0.074)	0.105 (0.108)	0.109 (0.105)	0.105 (0.107)	0.110 (0.107)
			0.22	0.105 (0.102)	0.101 (0.100)	0.071 (0.071)	0.105 (0.103)	0.102 (0.101)	0.105 (0.103)	0.102 (0.103)
			0.69	0.089 (0.091)	0.093 (0.089)	0.073 (0.071)	0.096 (0.097)	0.096 (0.096)	0.096 (0.097)	0.099 (0.097)
		0.5	0	0.089 (0.088)	0.084 (0.085)	0.060 (0.060)	0.089 (0.088)	0.085 (0.085)	0.089 (0.088)	0.087 (0.088)
			0.22	0.086 (0.088)	0.083 (0.085)	0.059 (0.061)	0.087 (0.089)	0.085 (0.086)	0.087 (0.089)	0.087 (0.089)
			0.69	0.089 (0.088)	0.087 (0.086)	0.063 (0.065)	0.092 (0.093)	0.092 (0.090)	0.093 (0.093)	0.093 (0.094)
	0.8	0.2	0	0.084 (0.087)	0.075 (0.076)	0.055 (0.057)	0.084 (0.087)	0.085 (0.085)	0.085 (0.086)	0.090 (0.090)
			0.22	0.083 (0.084)	0.073 (0.074)	0.056 (0.056)	0.085 (0.087)	0.088 (0.080)	0.085 (0.085)	0.093 (0.091)
			0.69	0.075 (0.077)	0.071 (0.072)	0.066 (0.065)	0.085 (0.088)	0.100 (0.077)	0.085 (0.085)	0.102 (0.097)
		0.5	0	0.066 (0.070)	0.063 (0.063)	0.045 (0.046)	0.066 (0.070)	0.070 (0.069)	0.066 (0.070)	0.072 (0.073)
			0.22	0.068 (0.068)	0.062 (0.062)	0.046 (0.046)	0.070 (0.070)	0.071 (0.069)	0.071 (0.070)	0.072 (0.073)
			0.69	0.059 (0.059)	0.057 (0.057)	0.048 (0.047)	0.069 (0.071)	0.072 (0.069)	0.069 (0.071)	0.072 (0.075)

[†] Analytical standard errors which are the average standard errors over 1000 simulations are summarized in the brackets.

1978 at the Centre for Prognosis Studies in Rheumatic Disease at the Toronto Western Hospital (Gladman and Chandran, 2011). Patients diagnosed with PsA were recruited and evaluated upon entry into the study with scheduled radiographic examinations every six months (Gladman et al, 1987). Assessments are conducted across 64 joints of the hands and feet. One way to determine the disease states is based on the number of damaged joints where a damaged joint is identified as deformed, ankylosed, flail, or surgically treated. We then consider four disease states represented by the multistate diagram in Figure 4.3 where state 0 denotes no joint damage, state 1 represents oligoarthritis with 1-4 damaged joints, state 2 represents polyarthritis with 5-9 damaged joints, and state 3 is entered upon the development of severe polyarthritis with more than 10 damaged joints.

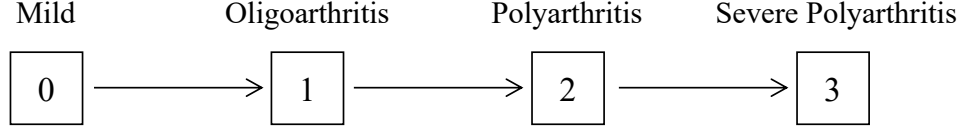


Figure 4.3: A progressive model for the development of damaged joints in patients with psoriatic arthritis.

Our interest is to investigate the risk of developing joint damage in patients with PsA in relation to the presence of the human leukocyte antigen B27 (HLA-B27) marker which has been linked to autoimmune conditions. Since testing for HLA markers through biosamples can be costly in large cohort studies, we employ a two-phase design to offer a more cost-effective data collection approach. We used a sample of 1,116 patients for analysis and consider the intensity-based regression modeling for the disease state transitions. The covariates included HLA-B27 status ($X_1 = 1$ if HLA-B27 is positive, $X_1 = 0$ otherwise), gender ($X_2 = 1$ for male, $X_2 = 0$ for female), age of PsA onset (in years) (X_3), and family history of PsA ($X_4 = 1$ if yes, $X_4 = 0$ otherwise). The transition intensity model takes the form of

$$\lambda_{k-1,k}(\mathbf{X}) = \lambda_{k-1,k} \exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4), \quad k = 1, 2, 3$$

for the disease process. We treat the HLA-B27 marker as the expensive covariate while gender, age of PsA onset, and family history of PsA are considered inexpensive covariates which are available for all individuals in the phase I sample. We select $n_2 = 480$ patients for HLA-B27 marker assessment using simple random sampling, balanced sampling, and score residual-dependent sampling. In balanced sampling strata are defined by the disease states at the last visit (states 0, 1, 2, 3), gender (male, female), age of PsA onset (≤ 40 , > 40), and family history of PsA (yes, no) which leads to 32 exclusive strata. We implemented the maximum likelihood, conditional maximum likelihood, and inverse probability weighting methods for estimating the effect of the HLA-B27 marker. The results along with those from a full data analysis are summarized in Table 4.3.

The likelihood-based estimating methods yields more efficient estimates than the inverse probability weighting approach. The regression coefficient estimates from the score residual-dependent phase II sub-sampling framework using the maximum likelihood method are very similar to those obtained from the full sample analysis, and Table 4.4 summarizes the estimates obtained from the maximum likelihood method with score residual-dependent sampling. The baseline transition intensity from state 2 to state 3 ($\lambda_{23} = 0.085$) is the highest which indicates that PsA patients in the early polyarthritis stage are at a higher

Table 4.3: Estimates with standard errors from maximum likelihood (ML), conditional maximum likelihood (CML), and inverse probability weighting (IPW) approaches for the two-phase designs with simple random sampling (SRS), balanced sampling (BAL), and residual-based sampling (RES) employed in the joint damage data of patients with psoriatic arthritis; phase I sample size $n = 1116$, and phase II sample size $n_2 = 480$.

Parameter	Full data	ML			CML		IPW	
		SRS	BAL	RES	SRS	BAL	SRS	BAL
λ_{01}	0.056 (0.007)	0.053 (0.006)	0.058 (0.006)	0.057 (0.006)	0.045 (0.008)	0.061 (0.010)	0.065 (0.016)	0.039 (0.010)
λ_{12}	0.054 (0.007)	0.051 (0.006)	0.057 (0.006)	0.055 (0.006)	0.044 (0.008)	0.073 (0.012)	0.065 (0.015)	0.039 (0.010)
λ_{23}	0.083 (0.011)	0.078 (0.009)	0.087 (0.010)	0.085 (0.010)	0.072 (0.012)	0.090 (0.014)	0.087 (0.023)	0.058 (0.016)
HLA-B27 (Positive vs. Negative)	0.277 (0.089)	0.491 (0.102)	0.199 (0.126)	0.300 (0.080)	0.288 (0.114)	0.147 (0.126)	0.188 (0.158)	0.342 (0.211)
Gender (Male vs. Female)	-0.084 (0.070)	-0.098 (0.064)	-0.094 (0.063)	-0.098 (0.063)	0.052 (0.101)	-0.196 (0.090)	-0.095 (0.117)	0.020 (0.137)
Age of PsA Onset	0.004 (0.003)	0.005 (0.003)	0.003 (0.002)	0.003 (0.002)	0.007 (0.004)	0.002 (0.003)	0.000 (0.005)	0.011 (0.005)
Family History of PsA (Yes vs. No)	0.089 (0.094)	0.080 (0.082)	0.088 (0.081)	0.105 (0.082)	0.001 (0.136)	0.073 (0.092)	0.067 (0.158)	0.087 (0.130)

risk of progressing to the next state which is the severe polyarthritis compare to patients with fewer damaged joints. Additionally PsA patients who are HLA-B27 positive are significantly associated with a higher risk of developing more damaged joints (hazard ratio (HR)=1.35). Although not statistically significant, patients with a family history of PsA also exhibited a higher risk of progressing to more damaged joints.

Table 4.4: Summary of estimates obtained from maximum likelihood (ML) with residual-dependent sampling (RES) employed in phase I; covariates include HLA-B27 status, gender, age of onset of psoriatic arthritis (years), and family history of psoriatic arthritis.

	Level	Est.	HR	95% CI of HR
λ_{01}		0.057		
λ_{12}		0.055		
λ_{23}		0.085		
HLA-B27	Positive vs. Negative	0.300	1.350	(1.153, 1.580)
Gender	Male vs. Female	-0.098	0.907	(0.802, 1.026)
Age of PsA Onset		0.003	1.003	(0.999, 1.008)
Family History of PsA	Yes vs. No	0.105	1.110	(0.945, 1.305)

4.7 Remarks

We have considered the design of two-phase studies aimed at selecting biospecimens for biomarker assays to measure biomarkers of interest and estimate their association with disease progression. Likelihood, conditional likelihood, and IPW approaches are developed with various sampling strategies. The presentation of the framework in terms of the progressive process is simplified. In many applications there will be progressive processes with more complex state spaces, e.g. multiple paths. In cases where mortality rates are high, the state space should expand to incorporate death. This changes the setting in that death is typically observed subject to right-censoring and this may be the case even when the transitions for the non-terminal states are not observed due to intermittent classification of individuals at clinic visits. The optimal design for studying disease progression can also be extended in a multiparameter setting where the goal is to choose a phase II subsample to estimate the effect of the expensive biomarker with different coefficients across the transitions.

Generalized raking can also be considered which is aimed at optimizing weights based on inverse probability weighted estimating functions when covariates are missing or mis-measured (Breslow et al, 2009). Calibration is typically carried out through the use of auxiliary variables that are available for all individuals in the sample. We let $\mathbf{Z}_{iG}$ be the auxiliary variable for individual i . The original weights $d_i = 1/\pi_i$ are modified to new weights $w_i = g_i d_i$ so that w_i and d_i are as close as possible (Breslow et al, 2009) such that the weighted sum of the auxiliary variable in phase II is the same as the total auxiliary variable in phase I, i.e.

$$\sum_{i=1}^n \mathbf{Z}_{iG} = \sum_{i=1}^n R_i w_i \mathbf{Z}_{iG}. \quad (4.16)$$

The weights adjustment g_i can be solved by minimizing $\sum_{i=1}^n R_i G(w_i, d_i)$ subject to the constraint in (4.16) where $G(w_i, d_i)$ is the distance function (Deville and Särndal, 1992). Examples of distance functions are discussed in Deville and Särndal (1992). We can consider Poisson type deviance $G(w_i, d_i) = w_i \log(w_i/d_i) - w_i + d_i$ as the distance function here which guarantees the weights are always positive, and the resulting estimator is known as generalized raking (GR) estimator (Breslow et al, 2009). The efficiency of the GR estimator depends on the auxiliary variable we used for calibration, and good choices of auxiliary variables are the ones that are highly correlated with the parameter of interest (Breslow et al, 2009; Chen and Lumley, 2022; Amorim et al, 2024). Let $h_i(\boldsymbol{\theta})$ be the influence function for $\boldsymbol{\theta}$ where

$$h_i(\boldsymbol{\theta}) = I_i(\boldsymbol{\theta})^{-1} \mathcal{S}_{i1}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}),$$

and $I_i(\boldsymbol{\theta})$ is the fisher information matrix of $\boldsymbol{\theta}$ for subject i which can be estimated using the first derivatives of the transition probabilities (Kalbfleisch and Lawless, 1985). Breslow et al (2009) show that the conditional expectation of the influence function given the phase I data (i.e. $\mathbb{E}\{h_i(\boldsymbol{\theta}) \mid \mathcal{D}_i, \mathbf{X}_2\}$) is the optimal auxiliary variable which minimizes the asymptotic variance of the calibration estimator. The optimal auxiliary variable is typically unavailable as the conditional expectation $\mathbb{E}\{h_i(\boldsymbol{\theta}) \mid \mathcal{D}_i, \mathbf{X}_2\}$ depends on the unknown parameter $\boldsymbol{\theta}$. In practice this will be approximated by the empirical value of $h_i(\hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ is estimated from the complete dataset with imputed values of X_{i1} for those who are not selected into phase II. The estimators are then obtained by solving

$$\sum_{i=1}^n U_{i2}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\psi}) = \sum_{i=1}^n R_i w_i S_{i1}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}) = 0 .$$

The asymptotic variance of $\hat{\boldsymbol{\theta}}$ can be approximated by treating the calibrated weights w_i as the sampling weights (Lumley, 2011), and we have (Amorim et al, 2024)

$$\sqrt{n} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right) \xrightarrow{d} \text{MVN} \left(0, \mathcal{A}_G^{-1} \mathcal{B}_G (\mathcal{A}_G^{-1})' \right) \quad \text{as } n \rightarrow \infty ,$$

where $\mathcal{A}_G = \mathbb{E}\{-\partial U_{i2}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\psi}) / \partial \boldsymbol{\theta}\}$, $\mathcal{B}_G = \mathbb{E}\{U_{i2}(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\psi}) U_{i2}'(\mathcal{D}_i \mid \mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\psi})\}$.

Chapter 5

Review and Future Work

This thesis presents new statistical methods that address challenging problems in designing and analyzing both observational studies and randomized controlled trials to assess exposure or intervention effects in complex settings. General remarks were made at the end of each chapter regarding the specific contributions and findings. Here we briefly comment on some areas of possible future work.

5.1 Confounding with Multivariate Exposures

In Chapter 2 we present regression and IPW methods for evaluating the effect of multiple exposures on the outcome of interest with different types (i.e. continuous, binary, and semi-continuous). Confounding is addressed using a generalized propensity score (GPS) which relies on modeling the joint distribution of multiple exposures. The key elements of the GPS are the linear functions of confounders needed to construct this joint distribution. An important insight is that when confounders influence the association of the exposure variables, the way they do so must be incorporated into propensity score modeling to maintain consistency in estimates and validity of inference regarding the interactions between exposures. Matching or stratification (Wu et al, 2022) can be developed in this setting based on the generalized propensity score for dealing with confounders. Stringer et al (2024) develop a framework for benchmark dose analysis based on the additive dose-response models where they focus on the univariate exposure. The generalized additive model (GAM) (Hastie and Tibshirani, 1995) can be considered for more flexible modeling of the dose-response effect with multiple exposures among the exposed in the two-stage approach.

5.2 Design of Plasma Donation Trials

The methodological development in Chapter 3 is specifically motivated by challenges in blood donation research where the goal is to show that a new device or algorithm used to extract blood from donors is safe. For this specific setting we use a binary response indicating the occurrence of a serious hypotensive adverse event wherein the probability of such an event is very rare – this along with the desire to estimate the relative rate of events, leads us to the use of a log link function. A working independence assumption means the dependence structure does not need to be specified but protection against misspecification of the variance function (implicit from the use of the Poisson approximation and log link) and serial dependence in responses is ensured by the use of a robust covariance matrix. While we formulate our method based on a binary response, other types of responses could be considered provided they are measurable soon after the donation.

The derivations of the joint distributions of successive test statistics and specification of alpha spending functions are exploited in Chapter 3 to ensure interim analyses and stopping rules are possible while ensuring control of the type I error rate. A unique feature of our derivation is that we recognize and address the fact that individuals recruited early during the course of the accrual period will typically be different than those recruited later during the accrual period - specifically in the donation context frequent donors will be encountered at donation centers sooner than donors who tend to donate less frequently. This creates a unique data structure whereby information is accrued much more rapidly early in the study compared to later stages when the patient accrual is nearing completion. This has significant implications on the optimal timing of interim analyses and trials aiming to plan analyses at roughly equal increments of information must ensure that data collection, checking and reporting is possible soon after the launch of the accrual period to ensure safety can be effectively monitored. The potential for early stopping means that trials may be shorter than non-sequential trials if treatment differences can be detected early.

The derivations we provide in Chapter 3 can also be exploited when interest lies in conditional power (Lachin, 2005) calculations. If there is a concern that sample size calculations were inadequate due to the misspecification of one or more parameters (e.g. the baseline event rate determined by β_0) then the joint distribution of the current and a future test statistic can be useful to calculate the probability that the trial at the planned or some later end of the study will lead to rejection of the null hypothesis. This may be done in the context of a fixed sample trial or a sequential trial, and can be used as a basis for sample size re-estimation (Mano et al, 2023) or stochastic curtailment (Xie et al, 2022) of a trial.

5.3 Two-Phase Designs with Complex Diseases

In Chapter 4 we develop two-phase designs with multistate processes under intermittent observation. We first note that the motivating study has led to some assumptions that simplify the setting we consider but the framework can be naturally adapted to deal with more complex settings. We assume that the expensive and inexpensive covariates have common effects across all transitions in the multistate process in Chapter 4. The methods can be extended to allow the inexpensive covariates to have different effects across the different transitions. This simplicity comes from the fact that the results follow as we have described them by simply adding a subscript k to β_2 to allow it to vary across the transitions. If the marker of interest has a different effect across the K transitions there may be two situations. In some settings interest may lie in the effect of X_1 on a particular transition, in which case the parameter of interest remains a scalar and the methods can be adapted with some refinement. In another setting there may be interest in the several effects of X_1 in which case the parameter of interest becomes a vector $\beta_1 = (\beta_{11}, \dots, \beta_{1K})'$ where the second subscript indicates the transition to which the marker effect corresponds. Two-phase designs can be developed to optimize the power of global tests of effects, or through sequential phase II sub-sampling as developed in Mao et al (2024).

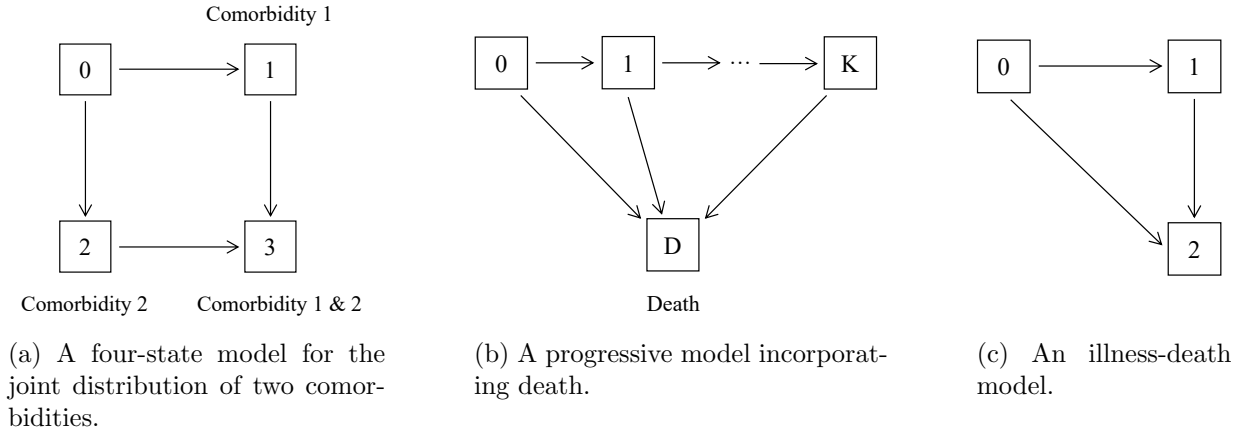


Figure 5.1: An illustration of three disease processes.

We consider a simple progressive process in Chapter 4, primarily because this is the setting of the motivating problem. In many applications there may be more general progressive processes in which more than one state can be entered from a given state. Figure 5.1(a)-(c) illustrate three general disease processes. We can consider multiple paths of the disease processes involving the development of different types of comorbidities (see Figure

5.1(a)) where the joint distribution of the comorbidities is needed. In settings where mortality rates are non-negligible one may add an absorbing state D corresponding to death, which can be entered from any one of states $0, 1, \dots, K$; see Figure 5.1(b). This also changes the setting somewhat by introducing a more complex observation scheme. In particular, even when transitions between the $K + 1$ disease states are not observable but the state occupied can be determined at intermittent clinic visits, death is typically observed subject only to right-censoring. A special case of this is the illness-death model (Xu et al, 2010) which arises if there is an initial healthy state, an intermediate state representing progression (i.e. as in progression or relapse in a cancer trial) and an absorbing state for death; see Figure 5.1(c). The dual observation schemes for the intermediate (intermittent) and terminal (right-censored) states represent a common structure often seen in cancer trials (Boruvka and Cook, 2016). If interest lies in identifying individuals to select for a phase II sub-sample to detect markers for progression, for those who were last observed in the initial state, if they die it will be unknown whether they passed through the intermediate state. This makes the value of sampling them in phase II unclear. Investigation of this setting based on large sample theory is warranted for each of the several approaches to the analysis of phase II data and design.

References

- Akkaya Hocagil T, Cook RJ, Jacobson SW, Jacobson JL, Ryan LM (2021) Propensity score analysis for a semi-continuous exposure variable: a study of gestational alcohol exposure and childhood cognition. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 184(4):1390–1413
- Albert M, Zhu Y, Moghekar A, Mori S, Miller MI, Soldan A, Pettigrew C, Selnes O, Li S, Wang MC (2018) Predicting progression from normal cognition to mild cognitive impairment for individuals at 5 years. *Brain* 141(3):877–887
- Amorim G, Tao R, Lotspeich S, Shaw PA, Lumley T, Patel RC, Shepherd BE (2024) Three-phase generalized raking and multiple imputation estimators to address error-prone data. *Statistics in Medicine* 43(2):379–394
- Andersen EB (1970) Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 32(2):283–301
- Andersen PK, Ravn H (2023) *Models for Multi-state Survival Data: Rates, Risks, and Pseudo-Values*. CRC Press
- Armitage P, McPherson C, Rowe B (1969) Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)* 132(2):235–244
- Austin PC (2019) Assessing covariate balance when using the generalized propensity score with quantitative or continuous exposures. *Statistical Methods in Medical Research* 28(5):1365–1377
- Bang H, Robins JM (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4):962–973
- Bebu I, Keshavarzi S, Gao X, Braffett BH, Canty AJ, Herman WH, Orchard TJ, Dagogo-Jack S, Nathan DM, Lachin JM, et al (2021) Genetic risk factors for CVD in type 1 diabetes: the DCCT/EDIC study. *Diabetes Care* 44(6):1309–1316

- Beelen R, Hoek G, van den Brandt PA, Goldbohm RA, Fischer P, Schouten LJ, Armstrong B, Brunekreef B (2008) Long-term exposure to traffic-related air pollution and lung cancer risk. *Epidemiology* 19(5):702–710
- Bégin P, Callum J, Jamula E, Cook R, Heddle NM, Tinmouth A, Zeller MP, Beaudoin-Bussi res G, Amorim L, Bazin R, et al (2021) Convalescent plasma for hospitalized patients with COVID-19: an open-label, randomized controlled trial. *Nature Medicine* 27(11):2012–2024
- Berger JO, Liseo B, Wolpert RL (1999) Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science* 14(1):1–28
- Berger M (2013) The value of incentives in blood donation. *Science* 341(6142):128–129
- Bobb JF, Claus Henn B, Valeri L, Coull BA (2018) Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. *Environmental Health* 17(1):1–10
- Boruvka A, Cook RJ (2016) Sieve estimation in a Markov illness-death process under dual censoring. *Biostatistics* 17(2):350–363
- Breslow N, Cain K (1988) Logistic regression for two-stage case-control data. *Biometrika* 75(1):11–20
- Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M (2009) Improved horvitz–thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Statistics in Biosciences* 1:32–49
- Brown J, Bakeman R, Coles C, Sexson W, Demi A (1998) Maternal drug use during pregnancy: Are preterm and full-term infants affected differently? *Developmental Psychology* 34(3):540–554
- Carmichael Olson H, Feldman J, Streissguth A, Sampson P, Bookstein F (1998) Neuropsychological deficits in adolescents with fetal alcohol syndrome. *Clinical and Experimental Research* 22(9):1998–2012
- Carter R, Jacobson J, Molteno C, Dodge N, Meintjes E, Jacobson S (2016) Fetal alcohol growth restriction and cognitive impairment. *Pediatrics* 138(2):e20160775
- Cazenave JP, Follea G, Bardiaux L, Boiron JM, Lafeuillade B, Debost M, Lioure B, Harousseau JL, Tabrizi R, Cahn JY, et al (2010) A randomized controlled clinical trial

- evaluating the performance and safety of platelets treated with mirasol pathogen reduction technology. *Transfusion* 50(11):2362–2375
- Chandler L, Richardson G, Gallagher J, Day N (1996) Prenatal exposure to alcohol and marijuana: Effects on motor development of preschool children. *Alcoholism: Clinical and Experimental Research* 20(3):455–461
- Chen HY, Li M (2011) Improving power and robustness for detecting genetic association with extreme-value sampling design. *Genetic Epidemiology* 35(8):823–830
- Chen J, Zhou Y (2023) Causal effect estimation for multivariate continuous treatments. *Biometrical Journal* 65(5):2200122
- Chen T, Lumley T (2022) Optimal sampling for design-based estimators of regression models. *Statistics in Medicine* 41(8):1482–1497
- Cochran WG (1968) The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 24(2):295–313
- Cole SR, Hernán MA (2008) Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* 168(6):656–664
- Coles C, Platzman K, Raskind-Hood C, Brown R, Falek A, Smith I (1997) A comparison of children affected by prenatal alcohol exposure and attention deficit, hyperactivity disorder. *Alcoholism: Clinical and Experimental Research* 21(1):150–161
- Collins R (2012) What makes uk biobank special? *The Lancet* 379(9822):1173–1174
- Collins RA, Wisniewski MK, Waters JH, Triulzi DJ, Yazer MH (2015) Effectiveness of multiple initiatives to reduce blood component wastage. *American Journal of Clinical Pathology* 143(3):329–335
- Cook RJ (1995) The design and analysis of randomized trials with recurrent events. *Statistics in Medicine* 14(19):2081–2098
- Cook RJ, Lawless JF (2007) *The Statistical Analysis of Recurrent Events*. Springer
- Cook RJ, Lawless JF (2018) *Multistate Models for the Analysis of Life History Data*. Chapman and Hall/CRC
- Cook RJ, Lawless JF (2021) Independence conditions and the analysis of life history studies with intermittent observation. *Biostatistics* 22(3):455–481

- Cook RJ, Wei W (2002) Selection effects in randomized trials with count data. *Statistics in Medicine* 21(4):515–531
- Cook RJ, Lee KA, Cuerden M, Cotton CA (2013) Inverse probability weighted estimating equations for randomized trials in transfusion medicine. *Statistics in Medicine* 32(25):4380–4399
- Covington ML, Voma C, Stowell SR (2023) The impact of the COVID-19 pandemic on source plasma donations. *Journal of Clinical Apheresis* 38(5):644–646
- Cox DR (2017) *The Theory of Stochastic Processes*. Routledge
- Cro S, Morris TP, Kenward MG, Carpenter JR (2020) Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: a practical guide. *Statistics in Medicine* 39(21):2815–2842
- Dale J, Paterson C, Tierney A, Ralston SH, Reid DM, Basu N, Harvie J, McKay ND, Saunders S, Wilson H, et al (2016) The scottish early rheumatoid arthritis (sera) study: an inception cohort and biobank. *BMC Musculoskeletal Disorders* 17:1–8
- Day NL, Helsel A, Sonon K, Goldschmidt L (2013) The association between prenatal alcohol exposure and behavior at 22 years of age. *Alcoholism: Clinical and Experimental Research* 37(7):1171–1178
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39(1):1–22
- Derkach A, Lawless JF, Sun L (2015) Score tests for association under response-dependent sampling designs for expensive covariates. *Biometrika* 102(4):988–994
- Devereux L, Watson PH, Mes-Masson AM, Luna-Crespo F, Thomas G, Pitman H, Speirs V, Hall AG, Bollinger N, Posada M, et al (2019) A review of international biobanks and networks: Success factors and key benchmarks—a 10-year retrospective review. *Biopreservation and Biobanking* 17(6):512–519
- Deville JC, Särndal CE (1992) Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87(418):376–382
- Di Credico G, Edefonti V, Polesel J, Pauli F, Torelli N, Serraino D, Negri E, Luce D, Stucker I, Matsuo K, et al (2019) Joint effects of intensity and duration of cigarette

- smoking on the risk of head and neck cancer: A bivariate spline model approach. *Oral oncology* 94:47–57
- Ding J, Zhou H, Liu Y, Cai J, Longnecker MP (2014) Estimating effect of environmental contaminants on women’s subfecundity for the moba study data with an outcome-dependent sampling scheme. *Biostatistics* 15(4):636–650
- Dominici F, Peng RD, Barr CD, Bell ML (2010) Protecting human health from air pollution: shifting from a single-pollutant to a multi-pollutant approach. *Epidemiology (Cambridge, Mass)* 21(2):187
- Duell EJ (2012) Epidemiology and potential mechanisms of tobacco smoking and heavy alcohol consumption in pancreatic cancer. *Molecular Carcinogenesis* 51(1):40–52
- Dummer TJ, Awadalla P, Boileau C, Craig C, Fortier I, Goel V, Hicks JM, Jacquemont S, Knoppers BM, Le N, et al (2018) The Canadian partnership for tomorrow project: a pan-Canadian platform for research on chronic disease prevention. *Canadian Medical Association Journal* 190(23):E710–E717
- Feng P, Zhou XH, Zou QM, Fan MY, Li XS (2012) Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine* 31(7):681–697
- Fitzmaurice GM, Laird NM, Ware JH (2012) *Applied Longitudinal Analysis*. John Wiley & Sons
- Fong C, Hazlett C, Imai K (2018) Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics* 12(1):156–177
- Geng Z, Guo J, Lau TS, Fung WK (2001) Confounding, homogeneity and collapsibility for causal effects in epidemiologic studies. *Statistica Sinica* 11(1):63–75
- Ghamdi AH (2020) Clinical predictors of diabetic retinopathy progression; a systematic review. *Current Diabetes Reviews* 16(3):242–247
- Gladman DD, Chandran V (2011) Observational cohort studies: lessons learnt from the University of Toronto Psoriatic Arthritis Program. *Rheumatology* 50(1):25–31
- Gladman DD, Shuckett R, Russell M, Thorne J, Schachter R (1987) Psoriatic arthritis (PSA)-an analysis of 220 patients. *QJM: An International Journal of Medicine* 62(2):127–141

- Glynn P, Greenland P (2020) Contributions of the uk biobank high impact papers in the era of precision medicine. *European Journal of Epidemiology* 35(1):5–10
- Goddard KA, Smith KS, Chen C, McMullen C, Johnson C (2009) Biobank recruitment: motivations for nonparticipation. *Biopreservation and Biobanking* 7(2):119–121
- Guidoum AC (2015) Kernel estimator and bandwidth selection for density and its derivatives. Department of Probabilities and Statistics, University of Science and Technology, Houari Boumediene, Algeria
- Hartmann J, Ragusa MJ, Burchardt ER, Manukyan Z, Popovsky MA, Leitman SF (2021) Personalized collection of plasma from healthy donors: a randomized controlled trial of a novel technology-enabled nomogram. *Transfusion* 61(6):1789–1798
- Hastie T, Tibshirani R (1995) Generalized additive models for medical research. *Statistical Methods in Medical Research* 4(3):187–196
- Haukoos JS, Lewis RJ (2015) The propensity score. *JAMA* 314(15):1637–1638
- Hedeker D, Gibbons RD (2006) *Longitudinal Data Analysis*. Wiley-Interscience
- Hernán MA, Robins JM (2006) Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health* 60(7):578–586
- Hernán MA, Robins JM (2010) *Causal inference: What If*. Boca Raton: Chapman & Hall/CRC
- Hirano K, Imbens GW (2004) *The Propensity Score with Continuous Treatments*, John Wiley & Sons, Ltd, chap 7, pp 73–84
- Hoyne H, May P, Kalberg W, Kodituwakku P, Gossage J, Trujillo P, Buckley D, Miller J, Aragon A, Khaole N, Viljoen D (2005) A practical clinical approach to diagnosis of fetal alcohol spectrum disorders: clarification of the 1996 institute of medicine criteria. *Pediatrics* 115(1):39–47
- Imai K, Van Dyk DA (2004) Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* 99(467):854–866
- Imbens GW (2000) The role of the propensity score in estimating dose-response functions. *Biometrika* 87(3):706–710

- Jackson CH (2011) Multi-state models for panel data: The msm package for R. *Journal of Statistical Software* 38(8):1–29, DOI 10.18637/jss.v038.i08
- Jacobson JL, Jacobson SW, Sokol RJ, Martier SS, Ager JW, Kaplan-Estrin MG (1993) Teratogenic effects of alcohol on infant development. *Alcoholism: Clinical and Experimental Research* 17(1):174–183
- Jacobson JL, Dodge NC, Burden MJ, Klorman R, Jacobson SW (2011) Number processing in adolescents with prenatal alcohol exposure and adhd: differences in the neurobehavioral phenotype. *Alcoholism: Clinical and Experimental Research* 35(3):431–442
- Jacobson JL, Akkaya-Hocagil T, Jacobson SW, Coles CD, Richardson GA, Olson HC, Day NL, Carter RC, Dodge NC, Dang KD, Ryan LM, Cook RJ (2024) A dose–response analysis of the effects of prenatal alcohol exposure on cognitive development. *Alcohol: Clinical and Experimental Research* 48(4):623–639
- Jacobson S, Carr L, Croxford J, Sokol R, Li T, Jacobson J (2006) Protective effects of the alcohol dehydrogenase-ADH1B allele in children exposed to alcohol during pregnancy. *The Journal of Pediatrics* 148(1):30–37
- Jacobson S, Stanton M, Molteno C, Burden M, Fuller D, Hoyme H, Robinson L, Khaole N, Jacobson J (2008) Impaired eyeblink conditioning in children with fetal alcohol syndrome. *Alcoholism: Clinical and Experimental Research* 32(2):365–372
- Jacobson SW, Chiodo LM, Sokol RJ, Jacobson JL (2002) Validity of maternal report of prenatal alcohol, cocaine, and smoking in relation to neurobehavioral outcome. *Pediatrics* 109(5):815–825
- Jacobson SW, Jacobson JL, Sokol RJ, Chiodo LM, Corobana R (2004) Maternal age, alcohol abuse history, and quality of parenting as moderators of the effects of prenatal alcohol exposure on 7.5-year intellectual function. *Alcoholism: Clinical and Experimental Research* 28(11):1732–1745
- James Hung H, Wang SJ, Tsong Y, Lawrence J, O’Neil RT (2003) Some fundamental issues with non-inferiority testing in active controlled trials. *Statistics in Medicine* 22(2):213–225
- Joe H, Xu JJ (1996) The estimation method of inference functions for margins for multivariate models. Technical Report 166, Department of Statistics, University of British Columbia, Vancouver

- Kalbfleisch J, Lawless JF (1985) The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* 80(392):863–871
- Kalema G, Molenberghs G, Kassahun W (2016) Second-order generalized estimating equations for correlated count data. *Computational Statistics* 31:749–770
- Kodituwakku P (2007) Defining the behavioral phenotype in children with fetal alcohol spectrum disorders: a review. *Neuroscience & Biobehavioral Reviews* 31(2):192–201
- Lachin JM (2005) A review of methods for futility stopping based on conditional power. *Statistics in Medicine* 24(18):2747–2764
- Laird NM, Ware JH (1982) Random-effects models for longitudinal data. *Biometrics* 38(4):963–974
- Lan KKG, DeMets DL (1983) Discrete sequential boundaries for clinical trials. *Biometrika* 70(3):659–663
- Lawless J (2018) Two-phase outcome-dependent studies for failure times and testing for effects of expensive covariates. *Lifetime Data Analysis* 24:28–44
- Lawless J, Kalbfleisch J, Wild C (1999) Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61(2):413–438
- Lawless JF (1987) Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* 15(3):209–225
- Lawless JF (2011) *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons
- Lee JW, Demets DL (1991) Sequential comparison of changes with repeated measurements data. *Journal of the American Statistical Association* 86(415):757–762
- Lee SJ, Kim K, Tsiatis AA (1996) Repeated significance testing in longitudinal clinical trials. *Biometrika* 83(4):779–789
- Lewis CE, Thomas KG, Dodge NC, Molteno CD, Meintjes EM, Jacobson JL, Jacobson SW (2015) Verbal learning and memory impairment in children with fetal alcohol spectrum disorders. *Alcoholism: Clinical and Experimental Research* 39(4):724–732
- Li D, Lewinger JP, Gauderman WJ, Murcray CE, Conti D (2011) Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genetic Epidemiology* 35(8):790–799

- Li F (2019) Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics* 13(4):2389–2415
- Li K, Cook RJ (2024b) The design of large-scale plasma donation trials. Submitted
- Li K, Cook RJ (2024c) Group sequential analysis of marked point processes: Plasma donation trials. Submitted
- Li K, Cook RJ (2024d) Two-phase design of biomarker studies with progressive disease processes under intermittent observation. Manuscript
- Li K, Akkaya-Hocagil T, Cook RJ, Ryan LM, Carter RC, Dang KD, Jacobson JL, Jacobson SW (2024a) Use of generalized propensity scores for assessing effects of multiple exposures. *Statistics in Biosciences* 16(2):347–376
- Li N, Pham T, Cheng C, McElfresh DC, Metcalf RA, Russell WA, Birch R, Yurkovich JT, Montemayor-Garcia C, Lane WJ, et al (2023) Blood demand forecasting and supply management: an analytical assessment of key studies utilizing novel computational techniques. *Transfusion Medicine Reviews* 37(4):150768
- Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. *Biometrika* 73(1):13–22
- Lin DY, Zeng D, Tang ZZ (2013) Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proceedings of the National Academy of Sciences* 110(30):12247–12252
- Lipsitz SR, Laird NM, Harrington DP (1991) Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika* 78(1):153–160
- Little RJ, Rubin DB (2019) *Statistical Analysis with Missing Data*. John Wiley & Sons
- Louis TA (1982) Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 44(2):226–233
- Lumley T (2011) *Complex Surveys: A Guide to Analysis Using R*. John Wiley & Sons
- Lunceford JK, Davidian M (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 23(19):2937–2960

- Mano H, Tanaka Y, Orihara S, Moriya J (2023) Application of sample size re-estimation in clinical trials: A systematic review. *Contemporary Clinical Trials Communications* 36:101210
- Mao F, Cook RJ (2023) Two-phase designs with current status data. *Statistics in Medicine* 42(8):1207–1232
- Mao F, Cheung LC, Cook RJ (2024) Two-phase designs with failure time processes subject to nonsusceptibility. *Biometrics* 80(1):ujad038
- Marschner S, Goodrich R (2011) Pathogen reduction technology treatment of platelets, plasma and whole blood using riboflavin and uv light. *Transfusion Medicine and Hemotherapy* 38(1):8–18
- May P, Gossage J, Kalberg W, Robinson L, Buckley D, Manning M, Hoyme H (2009) Prevalence and epidemiologic characteristics of FASD from various research methods with an emphasis on recent in-school studies. *Developmental Disabilities Research Reviews* 15(3):176–192
- May P, Blankenship J, Marais AS, Gossage J, Kalberg W, Barnard R, De Vries M, Robinson L, Adnams C, Buckley D, Manning M, Jones K, Parry C, Hoyme H, Seedat S (2013) Approaching the prevalence of the full spectrum of fetal alcohol spectrum disorders in a South African population-based study. *Alcoholism: Clinical and Experimental Research* 37(5):818–830
- McCullagh P (2019) *Generalized Linear Models*. Routledge
- McGee G, Perkins NJ, Mumford SL, Kioumourtzoglou MA, Weisskopf MG, Schildcrout JS, Coull BA, Schisterman EF, Haneuse S (2020) Methodological issues in population-based studies of multigenerational associations. *American Journal of Epidemiology* 189(12):1600–1609
- McIsaac M, Cook RJ (2017) Statistical methods for incomplete data: Some results on model misspecification. *Statistical Methods in Medical Research* 26(1):248–267
- Moraes T, Lefebvre D, Chooniedass R, Becker A, Brook J, Denburg J, HayGlass K, Hegele R, Kollmann T, Macri J, et al (2015) The Canadian healthy infant longitudinal development birth cohort study: biological samples and biobanking. *Paediatric and Perinatal Epidemiology* 29(1):84–92
- Nelsen RB (2007) *An introduction to copulas*. Springer Science & Business Media

- O'Brien PC, Fleming TR (1979) A multiple testing procedure for clinical trials. *Biometrics* 35(3):549–556
- Pascual J, Falk R, Piessens F, Prusinski A, Docekal P, Robert M, Ferrer P, Luria X, Segarra R, Zayas J (2000) Consistent efficacy and tolerability of almotriptan in the acute treatment of multiple migraine attacks: results of a large, randomized, double-blind, placebo-controlled study. *Cephalalgia* 20(6):588–596
- Peakman TC, Elliott P (2008) The UK biobank sample handling and storage validation studies. *International Journal of Epidemiology* 37(suppl_1):i2–i6
- Petersen ERB, Nielsen AA, Christensen H, Hansen T, Pedersen O, Christensen CK, Brandslund I (2016) Vejle diabetes biobank—a resource for studies of the etiologies of diabetes and its comorbidities. *Clinical Epidemiology* 8:393–413
- Pocock SJ (1977) Group sequential methods in the design and analysis of clinical trials. *Biometrika* 64(2):191–199
- Preisser JS, Qaqish BF (2014) A comparison of methods for simulating correlated binary variables with specified marginal means and correlations. *Journal of Statistical Computation and Simulation* 84(11):2441–2452
- Prentice RL (1988) Correlated binary regression with covariates specific to each binary observation. *Biometrics* 44(4):1033–1048
- Prentice RL, Zhao LP (1991) Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* 47(3):825–839
- Qaqish BF (2003) A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika* 90(2):455–463
- Rahdar M, Wang L, Hu G (2018) A tri-level optimization model for inventory control with uncertain demand and lead time. *International Journal of Production Economics* 195:96–105
- Rahman P, Gladman D, Cook R, Zhou Y, Young G, Salonen D (1998) Radiological assessment in psoriatic arthritis. *British Journal of Rheumatology* 37(7):760–765
- Raina PS, Wolfson C, Kirkland SA, Griffith LE, Oremus M, Patterson C, Tuokko H, Penning M, Balion CM, Hogan D, et al (2009) The Canadian longitudinal study on aging (clsa). *Canadian Journal on Aging/La Revue canadienne du vieillissement* 28(3):221–229

- Reboussin DM, DeMets DL, Kim K, Lan KG (2000) Computations for group sequential boundaries using the lan-demets spending function method. *Controlled Clinical Trials* 21(3):190–207
- Reijs BL, Teunissen CE, Goncharenko N, Betsou F, Blennow K, Baldeiras I, Brosseron F, Cavedo E, Fladby T, Froelich L, et al (2015) The central biobank and virtual biobank of biomarkapd: a resource for studies on neurodegenerative diseases. *Frontiers in Neurology* 6:216
- Richardson G, Goldschmidt L, Willford J (2008) The effects of prenatal cocaine use on infant development. *Neurotoxicology and Teratology* 30(2):96–106
- Robins JM (2000) Marginal structural models versus structural nested models as tools for causal inference. In: *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, Springer, pp 95–133
- Robins JM, Rotnitzky A, Zhao LP (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 90(429):106–121
- Rodeghiero F, Stasi R, Gernsheimer T, Michel M, Provan D, Arnold DM, Bussel JB, Cines DB, Chong BH, Cooper N, et al (2009) Standardization of terminology, definitions and outcome criteria in immune thrombocytopenic purpura of adults and children: report from an international working group. *Blood, The Journal of the American Society of Hematology* 113(11):2386–2393
- Rombach I, Jenkinson C, Gray AM, Murray DW, Rivero-Arias O (2018) Comparison of statistical approaches for analyzing incomplete longitudinal patient-reported outcome data in randomized controlled trials. *Patient Related Outcome Measures* 9:197–209
- Rosenbaum PR (1987) Model-based direct adjustment. *Journal of the American Statistical Association* 82(398):387–394
- Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55
- Rosenbaum PR, Rubin DB (1984) Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79(387):516–524

- Rosenbaum PR, Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39(1):33–38
- Rosselli M, MacNaughtan J, Jalan R, Pinzani M (2013) Beyond scoring: a modern interpretation of disease progression in chronic liver disease. *Gut* 62(9):1234–1241
- Rotnitzky A, Jewell NP (1990) Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* 77(3):485–497
- Rotnitzky A, Wypij D (1994) A note on the bias of estimators with missing data. *Biometrics* 50(4):1163–1170
- Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66(5):688
- Rubin DB (1980) Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association* 75(371):591–593
- Sanchez AM, Ameti DI, Schreiber GB, Thomson RA, Lo A, Bethel J, Williams AE, Study RED (2001) The potential impact of incentives on future blood donation behavior. *Transfusion* 41(2):172–178
- Satten GA (1999) Estimating the extent of tracking in interval-censored chain-of-events data. *Biometrics* 55(4):1228–1231
- Scharfstein DO, Rotnitzky A, Robins JM (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* 94(448):1096–1120
- Scott AJ, Wild CJ (2011) Fitting regression models with response-biased samples. *Canadian Journal of Statistics* 39(3):519–536
- Shu D, Han P, Hennessy S, Miano TA (2022) Robust causal inference of drug-drug interactions. *arXiv preprint arXiv:220300062*
- Slud E, Wei L (1982) Two-sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association* 77(380):862–868
- Snowden JM, Rose S, Mortimer KM (2011) Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology* 173(7):731–738

- Spiessens B, Lesaffre E, Verbeke G, Kim K, DeMets DL (2000) An overview of group sequential methods in longitudinal clinical trials. *Statistical Methods in Medical Research* 9(5):497–515
- Splawa-Neyman J, Dabrowska DM, Speed TP (1990) On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science* 5(4):465–472
- Stratton K, Howe C, Battaglia F (1996) Fetal Alcohol Syndrome: Diagnosis, Epidemiology, Prevention, and Treatment. The National Academies, Washington, D.C.
- Streissguth A (2007) Offspring effects of prenatal alcohol exposure from birth to 25 years: The Seattle prospective longitudinal study. *Journal of Clinical Psychology in Medical Settings* 14:81–101
- Stringer A, Akkaya Hocagil T, Cook RJ, Ryan LM, Jacobson SW, Jacobson JL (2024) Semi-parametric benchmark dose analysis with monotone additive models. *Biometrics* 80(3):ujae098
- Sun Z, Tao Y, Li S, Ferguson KK, Meeker JD, Park SK, Batterman SA, Mukherjee B (2013) Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. *Environmental Health* 12(1):1–19
- Suzuki T, Wakai K, Matsuo K, Hirose K, Ito H, Kuriki K, Sato S, Ueda R, Hasegawa Y, Tajima K (2006) Effect of dietary antioxidants and risk of oral, pharyngeal and laryngeal squamous cell carcinoma according to smoking and drinking habits. *Cancer Science* 97(8):760–767
- Tao R, Zeng D, Lin DY (2020) Optimal designs of two-phase studies. *Journal of the American Statistical Association* 115(532):1946–1959
- Vegetabile BG, Gillen DL, Stern HS (2020) Optimally balanced gaussian process propensity scores for estimating treatment effects. *Journal of the Royal Statistical Society Series A: Statistics in Society* 183(1):355–377
- Wand MP, Jones MC (1994) Kernel Smoothing. CRC press
- Webert K, Cook RJ, Sigouin CS, Rebulla P, Heddle NM (2006) The risk of bleeding in thrombocytopenic patients with acute myeloid leukemia. *Haematologica* 91(11):1530–1537

- Wei L, Su JQ, Lachin JM (1990) Interim analyses with repeated measurements in a sequential clinical trial. *Biometrika* 77(2):359–364
- White H (1982) Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society* 50(1):1–25
- Williams JR, Crespi CM (2020) Causal inference for multiple continuous exposures via the multivariate generalized propensity score. *arXiv preprint arXiv:200813767*
- Williamson EJ, Forbes A, Wolfe R (2012) Doubly robust estimators of causal exposure effects with missing data in the outcome, exposure or a confounder. *Statistics in Medicine* 31(30):4382–4400
- Wilson R, Allegra L, Huchon G, Izquierdo JL, Jones P, Schaberg T, Sagnier PP, Group MS, et al (2004) Short-term and long-term outcomes of moxifloxacin compared to standard antibiotic treatment in acute exacerbations of chronic bronchitis. *CHEST* 125(3):953–964
- Wu X, Mealli F, Kioumourtzoglou MA, Dominici F, Braun D (2022) Matching on generalized propensity scores with continuous exposures. *Journal of the American Statistical Association* 119(545):1–29
- Xiang RF, Quinn JG, Watson S, Kumar-Misir A, Cheng C (2021) Application of unsupervised machine learning to identify areas of blood product wastage in transfusion medicine. *Vox Sanguinis* 116(9):955–964
- Xie T, Zhang P, Shih WJ, Tu Y, Lan KG (2022) Dynamic monitoring of ongoing clinical trials. *Statistics in Biopharmaceutical Research* 14(4):580–591
- Xu J, Kalbfleisch JD, Tai B (2010) Statistical analysis of illness–death processes and semi-competing risks data. *Biometrics* 66(3):716–725
- Yang C, Diao L, Cook RJ (2022) Adaptive response-dependent two-phase designs: Some results on robustness and efficiency. *Statistics in Medicine* 41(22):4403–4425
- Yule GU (1912) On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society* 75(6):579–652
- Zhao LP, Prentice RL (1990) Correlated binary regression using a quadratic exponential model. *Biometrika* 77(3):642–648

- Zhao P, Dong D, Dong R, Zhou Y, Hong Y, Xiao G, Li Z, Su X, Zheng X, Liu X, et al (2023) Development and validation of a nomogram for predicting the risk of vasovagal reactions after plasma donation. *Journal of Clinical Apheresis* 38(5):622–631
- Zhao S, van Dyk DA, Imai K (2020) Propensity score-based methods for causal inference in observational studies with non-binary treatments. *Statistical Methods in Medical Research* 29(3):709–727
- Zhong Y, Kennedy EH, Bodnar LM, Naimi AI (2021) AIPW: An R package for augmented inverse probability-weighted estimation of average causal effects. *American Journal of Epidemiology* 190(12):2690–2699
- Zhou Q, Cai J, Zhou H (2020) Semiparametric inference for a two-stage outcome-dependent sampling design with interval-censored failure time data. *Lifetime Data Analysis* 26:85–108

APPENDICES

Appendix A

Appendix to Chapter 2

A.1 Impact of a Misspecified Generalized Propensity Score

Let $\mathbf{A} = (A_1, A_2)'$, $B_1 = A_1 A_2$, and $\mathbf{X} = (\mathbf{A}', B_1)'$ contain the exposure data of interest, and let $\mathbf{Z} = (Z_1, Z_2, \dots, Z_r)'$ be an $r \times 1$ vector of confounding variables. The data generating model in (2.1) has the form

$$\mathbb{E}(Y \mid \mathbf{X}, \mathbf{Z}) = \mathbf{X}'\boldsymbol{\theta} + \mathbf{Z}'\boldsymbol{\zeta}, \quad (\text{A.1})$$

where $\boldsymbol{\theta} = (\alpha_0, \alpha_1, \alpha_2, \beta_1)'$, $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_r)'$. The regression model with the misspecified generalized propensity score $\tilde{\mathbf{S}}$ has the form

$$\mathbb{E}(Y \mid \mathbf{X}, \tilde{\mathbf{S}}) = \mathbf{X}'\boldsymbol{\varphi} + \tilde{\mathbf{S}}'\boldsymbol{\psi}, \quad (\text{A.2})$$

where $\mathbf{X} = (1, A_1, A_2, B_1)'$ and $\boldsymbol{\varphi} = (\varphi_0, \varphi_1, \varphi_2, \varphi_3)'$. The contribution to the least squares equation for $\boldsymbol{\varphi}$ from a single individual based on (A.2) is

$$\mathbf{U}(Y \mid \mathbf{X}, \tilde{\mathbf{S}}; \boldsymbol{\varphi}, \boldsymbol{\psi}) = \mathbf{V}(Y - \mathbf{X}'\boldsymbol{\varphi} - \tilde{\mathbf{S}}'\boldsymbol{\psi}), \quad (\text{A.3})$$

where $\mathbf{V} = (\mathbf{X}', \tilde{\mathbf{S}}')'$.

We first take the expectation of (A.3) with respect to $Y \mid \mathbf{X}, \mathbf{Z}$, to get

$$\mathbb{E}_{Y \mid \mathbf{X}, \mathbf{Z}}\{\mathbf{U}(Y \mid \mathbf{X}, \tilde{\mathbf{S}}; \boldsymbol{\varphi}, \boldsymbol{\psi})\} = \mathbf{V}(\mathbf{X}'\boldsymbol{\theta} + \mathbf{Z}'\boldsymbol{\zeta} - \mathbf{X}'\boldsymbol{\varphi} - \tilde{\mathbf{S}}'\boldsymbol{\psi}). \quad (\text{A.4})$$

Next we take the expectation of (A.4) with respect to $\mathbf{X} \mid \mathbf{Z}$ which gives the form of a $(4 + s) \times 1$ vector, where $s = \dim(\tilde{\mathbf{S}})$, with the first 4 elements of the vector as follows,

$$\begin{aligned}
& \alpha_0 + \alpha_1 \mathbb{E}(A_1 \mid \mathbf{Z}) + \alpha_2 \mathbb{E}(A_2 \mid \mathbf{Z}) + \beta_1 \mathbb{E}(B_1 \mid \mathbf{Z}) + \mathbf{Z}' \boldsymbol{\zeta} \\
& - (\varphi_0 + \varphi_1 \mathbb{E}(A_1 \mid \mathbf{Z}) + \varphi_2 \mathbb{E}(A_2 \mid \mathbf{Z}) + \varphi_3 \mathbb{E}(B_1 \mid \mathbf{Z}) + \tilde{\mathbf{S}}' \boldsymbol{\psi}) = 0 , \\
& \alpha_0 \mathbb{E}(A_1 \mid \mathbf{Z}) + \alpha_1 \mathbb{E}(A_1^2 \mid \mathbf{Z}) + \alpha_2 \mathbb{E}(B_1 \mid \mathbf{Z}) + \beta_1 \mathbb{E}(A_1 B_1 \mid \mathbf{Z}) + \mathbf{Z}' \boldsymbol{\zeta} \mathbb{E}(A_1 \mid \mathbf{Z}) \\
& - (\varphi_0 \mathbb{E}(A_1 \mid \mathbf{Z}) + \varphi_1 \mathbb{E}(A_1^2 \mid \mathbf{Z}) + \varphi_2 \mathbb{E}(B_1 \mid \mathbf{Z}) + \varphi_3 \mathbb{E}(A_1 B_1 \mid \mathbf{Z}) + \tilde{\mathbf{S}}' \boldsymbol{\psi} \mathbb{E}(A_1 \mid \mathbf{Z})) = 0 , \\
& \alpha_0 \mathbb{E}(A_2 \mid \mathbf{Z}) + \alpha_1 \mathbb{E}(B_1 \mid \mathbf{Z}) + \alpha_2 \mathbb{E}(A_2^2 \mid \mathbf{Z}) + \beta_1 \mathbb{E}(A_2 B_1 \mid \mathbf{Z}) + \mathbf{Z}' \boldsymbol{\zeta} \mathbb{E}(A_2 \mid \mathbf{Z}) \\
& - (\varphi_0 \mathbb{E}(A_2 \mid \mathbf{Z}) + \varphi_1 \mathbb{E}(B_1 \mid \mathbf{Z}) + \varphi_2 \mathbb{E}(A_2^2 \mid \mathbf{Z}) + \varphi_3 \mathbb{E}(A_2 B_1 \mid \mathbf{Z}) + \tilde{\mathbf{S}}' \boldsymbol{\psi} \mathbb{E}(A_2 \mid \mathbf{Z})) = 0 , \\
& \alpha_0 \mathbb{E}(B_1 \mid \mathbf{Z}) + \alpha_1 \mathbb{E}(A_1 B_1 \mid \mathbf{Z}) + \alpha_2 \mathbb{E}(A_2 B_1 \mid \mathbf{Z}) + \beta_1 \mathbb{E}(B_1^2 \mid \mathbf{Z}) + \mathbf{Z}' \boldsymbol{\zeta} \mathbb{E}(B_1 \mid \mathbf{Z}) \\
& - (\varphi_0 \mathbb{E}(B_1 \mid \mathbf{Z}) + \varphi_1 \mathbb{E}(A_1 B_1 \mid \mathbf{Z}) + \varphi_2 \mathbb{E}(A_2 B_1 \mid \mathbf{Z}) + \varphi_3 \mathbb{E}(B_1^2 \mid \mathbf{Z}) + \tilde{\mathbf{S}}' \boldsymbol{\psi} \mathbb{E}(B_1 \mid \mathbf{Z})) = 0 .
\end{aligned}$$

The expectation of the above four equations with respect to $\mathbf{Z}$ gives the following four equations,

$$\begin{aligned}
& \alpha_0 + \alpha_1 \mathbb{E}(A_1) + \alpha_2 \mathbb{E}(A_2) + \beta_1 \mathbb{E}(B_1) + \boldsymbol{\zeta}' \mathbb{E}(\mathbf{Z}) \\
& - (\varphi_0 + \varphi_1 \mathbb{E}(A_1) + \varphi_2 \mathbb{E}(A_2) + \varphi_3 \mathbb{E}(B_1) + \mathbb{E}(\tilde{\mathbf{S}}') \boldsymbol{\psi}) = 0 , \quad (\text{A.5})
\end{aligned}$$

$$\begin{aligned}
& \alpha_0 \mathbb{E}(A_1) + \alpha_1 \mathbb{E}(A_1^2) + \alpha_2 \mathbb{E}(A_1 A_2) + \beta_1 \mathbb{E}(A_1 B_1) + \boldsymbol{\zeta}' \{\mathbb{E}(A_1) \mathbb{E}(\mathbf{Z}) + \Sigma_1\} \\
& - (\varphi_0 \mathbb{E}(A_1) + \varphi_1 \mathbb{E}(A_1^2) + \varphi_2 \mathbb{E}(A_1 A_2) + \varphi_3 \mathbb{E}(A_1 B_1) + \boldsymbol{\psi}' \{\mathbb{E}(A_1) \mathbb{E}(\tilde{\mathbf{S}}) + \Lambda_1\}) = 0 , \quad (\text{A.6})
\end{aligned}$$

$$\begin{aligned}
& \alpha_0 \mathbb{E}(A_2) + \alpha_1 \mathbb{E}(A_1 A_2) + \alpha_2 \mathbb{E}(A_2^2) + \beta_1 \mathbb{E}(A_2 B_1) + \boldsymbol{\zeta}' \{\mathbb{E}(A_2) \mathbb{E}(\mathbf{Z}) + \Sigma_2\} \\
& - (\varphi_0 \mathbb{E}(A_2) + \varphi_1 \mathbb{E}(A_1 A_2) + \varphi_2 \mathbb{E}(A_2^2) + \varphi_3 \mathbb{E}(A_2 B_1) + \boldsymbol{\psi}' \{\mathbb{E}(A_2) \mathbb{E}(\tilde{\mathbf{S}}) + \Lambda_2\}) = 0 , \quad (\text{A.7})
\end{aligned}$$

$$\begin{aligned} & \alpha_0 \mathbb{E}(B_1) + \alpha_1 \mathbb{E}(A_1 B_1) + \alpha_2 \mathbb{E}(A_2 B_1) + \beta_1 \mathbb{E}(B_1^2) + \boldsymbol{\zeta}' \{ \mathbb{E}(B_1) \mathbb{E}(\mathbf{Z}) + \Sigma_3 \} \\ & - (\varphi_0 \mathbb{E}(B_1) + \varphi_1 \mathbb{E}(A_1 B_1) + \varphi_2 \mathbb{E}(A_2 B_1) + \varphi_3 \mathbb{E}(B_1^2) + \boldsymbol{\psi}' \{ \mathbb{E}(B_1) \mathbb{E}(\tilde{\mathbf{S}}) + \Lambda_3 \}) = 0, \end{aligned} \quad (\text{A.8})$$

where $\Sigma_k = \text{cov}(\mathbb{E}(A_k | \mathbf{Z}), \mathbf{Z})$, $\Sigma_3 = \text{cov}(\mathbb{E}(B_1 | \mathbf{Z}), \mathbf{Z})$, $\Lambda_k = \text{cov}(\mathbb{E}(A_k | \mathbf{Z}), \tilde{\mathbf{S}})$, and $\Lambda_3 = \text{cov}(\mathbb{E}(B_1 | \mathbf{Z}), \tilde{\mathbf{S}})$, $k = 1, 2$. Note that Σ_1, Σ_2 and Σ_3 are the $r \times 1$ vectors, where $r = \dim(\mathbf{Z})$; Λ_1, Λ_2 and Λ_3 are the $s \times 1$ vectors, where $s = \dim(\tilde{\mathbf{S}})$. Solving for φ_0 from (A.5), gives

$$\begin{aligned} \varphi_0 = & \alpha_0 + \alpha_1 \mathbb{E}(A_1) + \alpha_2 \mathbb{E}(A_2) + \beta_1 \mathbb{E}(B_1) + \boldsymbol{\zeta}' \mathbb{E}(\mathbf{Z}) \\ & - \varphi_1 \mathbb{E}(A_1) - \varphi_2 \mathbb{E}(A_2) - \varphi_3 \mathbb{E}(B_1) - \mathbb{E}(\tilde{\mathbf{S}}') \boldsymbol{\psi}. \end{aligned} \quad (\text{A.9})$$

Next replacing φ_0 in (A.6), (A.7), and (A.8) respectively, gives

$$\begin{aligned} & \alpha_1 \text{var}(A_1) + \alpha_2 \text{cov}(A_1, A_2) + \beta_1 \text{cov}(A_1, B_1) + \boldsymbol{\zeta}' \Sigma_1 \\ & - \varphi_1 \text{var}(A_1) - \varphi_2 \text{cov}(A_1, A_2) - \varphi_3 \text{cov}(A_1, B_1) - \boldsymbol{\psi}' \Lambda_1 = 0, \end{aligned} \quad (\text{A.10})$$

$$\begin{aligned} & \alpha_1 \text{cov}(A_1, A_2) + \alpha_2 \text{var}(A_2) + \beta_1 \text{cov}(A_2, B_1) + \boldsymbol{\zeta}' \Sigma_2 \\ & - \varphi_1 \text{cov}(A_1, A_2) - \varphi_2 \text{var}(A_2) - \varphi_3 \text{cov}(A_2, B_1) - \boldsymbol{\psi}' \Lambda_2 = 0, \end{aligned} \quad (\text{A.11})$$

$$\begin{aligned} & \alpha_1 \text{cov}(A_1, B_1) + \alpha_2 \text{cov}(A_1, B_1) + \beta_1 \text{var}(B_1) + \boldsymbol{\zeta}' \Sigma_3 \\ & - \varphi_1 \text{cov}(A_1, B_1) - \varphi_2 \text{cov}(A_2, B_1) - \varphi_3 \text{var}(B_1) - \boldsymbol{\psi}' \Lambda_3 = 0. \end{aligned} \quad (\text{A.12})$$

Solving for φ_3 from (A.12) we obtain

$$\begin{aligned} \varphi_3 = & \beta_1 + \alpha_1 \frac{\text{cov}(A_1, B_1)}{\text{var}(B_1)} + \alpha_2 \frac{\text{cov}(A_2, B_1)}{\text{var}(B_1)} + \boldsymbol{\zeta}' \frac{\Sigma_3}{\text{var}(B_1)} \\ & - \varphi_1 \frac{\text{cov}(A_1, B_1)}{\text{var}(B_1)} - \varphi_2 \frac{\text{cov}(A_2, B_1)}{\text{var}(B_1)} - \boldsymbol{\psi}' \frac{\Lambda_3}{\text{var}(B_1)}. \end{aligned} \quad (\text{A.13})$$

Then plugging φ_3 into (A.10) and (A.11), we can solve for φ_1 and φ_2 as

$$\begin{aligned}\tilde{\varphi}_1 &= \alpha_1 + \zeta' \Delta_{11} + \psi' \Delta_{21} , \\ \tilde{\varphi}_2 &= \alpha_2 + \zeta' \Delta_{12} + \psi' \Delta_{22} ,\end{aligned}$$

where

$$\begin{aligned}\Delta_{1k} &= \frac{(\rho_{12} - \rho_{13}\rho_{23})(\frac{\sigma_k^2 \Sigma_{3-k}}{\sigma_1 \sigma_2} - \frac{\sigma_k}{\sigma_3} \frac{\rho_{13}\rho_{23}}{\rho_{k3}} \Sigma_3) + (\frac{\rho_{13}^2 \rho_{23}^2}{\rho_{k3}^2} - 1)(\Sigma_k - \frac{\sigma_k}{\sigma_3} \rho_{k3} \Sigma_3)}{\sigma_k^2 \{(\rho_{12} - \rho_{13}\rho_{23})^2 - (1 - \rho_{13}^2)(1 - \rho_{23}^2)\}} , \\ \Delta_{2k} &= \frac{(\rho_{12} - \rho_{13}\rho_{23})(\frac{\sigma_k^2 \Lambda_{3-k}}{\sigma_1 \sigma_2} - \frac{\sigma_k}{\sigma_3} \frac{\rho_{13}\rho_{23}}{\rho_{k3}} \Lambda_3) + (\frac{\rho_{13}^2 \rho_{23}^2}{\rho_{k3}^2} - 1)(\Lambda_k - \frac{\sigma_k}{\sigma_3} \rho_{k3} \Lambda_3)}{\sigma_k^2 \{(\rho_{12} - \rho_{13}\rho_{23})^2 - (1 - \rho_{13}^2)(1 - \rho_{23}^2)\}} ,\end{aligned}$$

$\sigma_k^2 = \text{var}(A_k)$, $\rho_{k3} = \text{corr}(A_k, B_1)$, $k = 1, 2$, $\sigma_3^2 = \text{var}(B_1)$, and $\rho_{12} = \text{corr}(A_1, A_2)$. Replacing $\tilde{\varphi}_1$ and $\tilde{\varphi}_2$ in (A.13) gives

$$\tilde{\varphi}_3 = \beta_1 + \zeta' \left(\frac{\Sigma_3}{\sigma_3^2} - \Delta_{11} \frac{\rho_{13}\sigma_1}{\sigma_3} - \Delta_{12} \frac{\rho_{23}\sigma_2}{\sigma_3} \right) - \psi' \left(\frac{\Lambda_3}{\sigma_3^2} + \Delta_{21} \frac{\rho_{13}\sigma_1}{\sigma_3} + \Delta_{22} \frac{\rho_{23}\sigma_2}{\sigma_3} \right) .$$

A.2 Second-Order Generalized Estimating Equations (GEE2)

Let $\bar{\gamma} = (\gamma'_0, \gamma'_1, \gamma'_2, \gamma'_3)'$ with γ_j a $g_j \times 1$ vector, $j = 0, 1, 2, 3$. Furthermore let $\gamma = (\gamma'_1, \gamma'_2, \gamma'_3)'$ and $\Gamma = (\gamma'_1, \sigma_1^2, \gamma'_2, \sigma_2^2, \gamma'_3)'$ be a $q \times 1$ vector with $q = \sum_{j=1}^3 g_j + 2$. For a sample of size n , we have an inverse density weighted estimating function for φ with stabilized weights w denoted as

$$U_0(Y \mid \mathbf{X}; \varphi, \Gamma, \Gamma^\dagger, \gamma_0, \gamma_0^\dagger) = \sum_{i=1}^n U_{i0}(Y_i \mid \mathbf{X}_i; \varphi, \Gamma, \Gamma^\dagger, \gamma_0, \gamma_0^\dagger) , \quad (\text{A.14})$$

where

$$U_{i0}(Y_i \mid \mathbf{X}_i; \varphi, \Gamma, \Gamma^\dagger, \gamma_0, \gamma_0^\dagger) = w_i^s(\mathbf{X}_i, \mathbf{Z}_i) D(\mathbf{X}_i) (Y_i - \mathbf{X}_i' \varphi) ,$$

and

$$w_i^s(\mathbf{X}_i, \mathbf{Z}_i) = \frac{\mathbb{I}(\mathbf{X}_i = \mathbf{x}_i) \{ \pi_{i2}(\gamma_0^\dagger) f(l_{i1}, l_{i2} \mid A_i^+ = 1; \Gamma^\dagger) \}^{A_i^+} \{ 1 - \pi_{i2}(\gamma_0^\dagger) \}^{1-A_i^+}}{\{ \pi_{i1}(\mathbf{Z}_{i0}; \gamma_0) f(l_{i1}, l_{i2} \mid A_i^+ = 1, \mathbf{Z}_i; \Gamma_i) \}^{A_i^+} \{ 1 - \pi_{i1}(\mathbf{Z}_{i0}; \gamma_0) \}^{1-A_i^+}} ,$$

and $D(\mathbf{X}_i) = (1, A_i^+, A_i^+ L_{i1}, A_i^+ L_{i2}, A_i^+ L_{i1} L_{i2})'$. For the numerator, we consider

$$L_1, L_2 \mid A^+ = 1 \sim BVN \left(\begin{pmatrix} \gamma_4 \\ \gamma_5 \end{pmatrix}, \begin{pmatrix} \sigma_3^2 & \sigma_3 \sigma_4 \rho_2 \\ \sigma_3 \sigma_4 \rho_2 & \sigma_4^2 \end{pmatrix} \right),$$

where $(\hat{\gamma}_4, \hat{\gamma}_5)' = (\bar{l}_1, \bar{l}_2)'$, $\hat{\sigma}_{k+2}^2 = \frac{1}{n-1} \sum_{i=1}^n (l_{ik} - \bar{l}_k)^2$, $k = 1, 2$, $\hat{\rho}_2 = \text{corr}(l_1, l_2)$, and let $\mathbf{\Gamma}^\dagger = (\gamma_4, \gamma_5, \sigma_3^2, \sigma_4^2, \rho_2)'$.

The set of unknown parameters $\mathbf{\Gamma}$ and γ_0 is estimated through the second-order generalized estimating equations (GEE2) (Zhao and Prentice, 1990; Kalema et al, 2016) for the indicator A^+ , the exposure model of L_1, L_2 given $A^+ = 1$ and $\mathbf{Z}$. We write the estimating functions as

$$U_{11}(L_1, L_2 \mid A^+ = 1, \mathbf{Z}; \mathbf{\Gamma}) = \sum_{i=1}^n U_{i11}(L_{i1}, L_{i2} \mid A_i^+ = 1, \mathbf{Z}_i; \mathbf{\Gamma}), \quad (\text{A.15})$$

$$U_{12}(A^+ \mid \mathbf{Z}_0; \gamma_0) = \sum_{i=1}^n U_{i12}(A_i^+ \mid \mathbf{Z}_{i0}; \gamma_0), \quad (\text{A.16})$$

where

$$U_{i11}(L_{i1}, L_{i2} \mid A_i^+ = 1, \mathbf{Z}_i; \mathbf{\Gamma}) = A_i^+ D_i' \Sigma_i^{-1} (W_i - \Omega_i),$$

and

$$U_{i12}(A_i^+ \mid \mathbf{Z}_{i0}; \gamma_0) = \bar{Z}_{i0} \{A_i^+ - \pi_{i1}(\mathbf{Z}_{i0}; \gamma_0)\},$$

with $W_i = (L_{i1}, L_{i1}^2, L_{i2}, L_{i2}^2, B_i)'$, $S_k(\mathbf{Z}_k; \gamma_k) = \mathbf{Z}_k' \gamma_k$, $k = 1, 2$ and

$$\Omega_i = \mathbb{E}(W_i \mid A_i^+ = 1, \mathbf{Z}_i; \mathbf{\Gamma}) = \begin{pmatrix} S_1(\mathbf{Z}_{i1}; \gamma_1) \\ \sigma_1^2 + S_1^2(\mathbf{Z}_{i1}; \gamma_1) \\ S_2(\mathbf{Z}_{i2}; \gamma_2) \\ \sigma_2^2 + S_2^2(\mathbf{Z}_{i2}; \gamma_2) \\ \mathbb{E}(B_i \mid A_i^+ = 1, \mathbf{Z}_i; \mathbf{\Gamma}) \end{pmatrix},$$

$$D_i = \frac{\partial \Omega_i}{\partial \mathbf{\Gamma}'} = \begin{pmatrix} \mathbf{Z}_{i1}' & 0 & \mathbf{0} & 0 & \mathbf{0} \\ \frac{\partial S_1^2(\mathbf{Z}_{i1}; \gamma_1)}{\partial \gamma_1'} & 1 & \mathbf{0} & 0 & \mathbf{0} \\ \mathbf{0} & 0 & \mathbf{Z}_{i2}' & 0 & \mathbf{0} \\ \mathbf{0} & 0 & \frac{\partial S_2^2(\mathbf{Z}_{i2}; \gamma_2)}{\partial \gamma_2'} & 1 & \mathbf{0} \\ S_2(\mathbf{Z}_{i2}; \gamma_2) \mathbf{Z}_{i1}' & \rho_1(\mathbf{Z}_{i3}; \gamma_3) \frac{\sigma_2}{2\sigma_1} & S_1(\mathbf{Z}_{i1}; \gamma_1) \mathbf{Z}_{i2}' & \rho_1(\mathbf{Z}_{i3}; \gamma_3) \frac{\sigma_1}{2\sigma_2} & \sigma_1 \sigma_2 \frac{\partial \rho_1(\mathbf{Z}_{i3}; \gamma_3)}{\partial \gamma_3'} \end{pmatrix}$$

is a $5 \times q$ matrix, where

$$\begin{aligned}\mathbb{E}(B_i \mid A_i^+ = 1, \mathbf{Z}_i; \mathbf{\Gamma}) &= \rho_1(\mathbf{Z}_{i3}; \gamma_3) \sigma_1 \sigma_2 + S_1(\mathbf{Z}_{i1}; \gamma_1) S_2(\mathbf{Z}_{i2}; \gamma_2) , \\ \frac{\partial S_k^2(\mathbf{Z}_{ik}; \gamma_k)}{\partial \gamma'_k} &= 2(\mathbf{Z}'_{ik} \gamma_{ik}) \mathbf{Z}'_{ik}, \quad k = 1, 2 , \\ \frac{\partial \rho_1(\mathbf{Z}_{i3}; \gamma_3)}{\partial \gamma'_3} &= 2 \exp(\mathbf{Z}'_3 \gamma_3) \mathbf{Z}'_{i3} / (\exp(\mathbf{Z}'_3 \gamma_3) + 1)^2 .\end{aligned}$$

Additionally, $\Sigma_i = \text{cov}(W_i \mid \mathbf{Z}_i)$, but we can simplify it and use a partial working independence structure with $\Sigma_i = \text{cov}(W_i \mid \mathbf{Z}_i; \mathbf{\Gamma})$ specified as,

$$\begin{pmatrix} \sigma_1^2 & & \rho_1(\mathbf{Z}_{i3}) \sigma_1 \sigma_2 & & \\ & \text{var}(L_{i1}^2 \mid A_i^+ = 1, \mathbf{Z}_i) & & & \\ \rho_1(\mathbf{Z}_{i3}) \sigma_1 \sigma_2 & & \sigma_2^2 & & \\ & & & \text{var}(L_{i2}^2 \mid A_i^+ = 1, \mathbf{Z}_i) & \\ & & & & \text{var}(B_i \mid A_i^+ = 1, \mathbf{Z}_i) \end{pmatrix}.$$

Using moment generating functions, we can obtain higher order moments in the matrix as

$$\begin{aligned}\text{var}(L_{i1}^2 \mid A_i^+ = 1, \mathbf{Z}_i) &= 4S_1(\mathbf{Z}_{i1}; \gamma_1) \sigma_1^2 + 2\sigma_1^4 , \\ \text{var}(L_{i2}^2 \mid A_i^+ = 1, \mathbf{Z}_i) &= 4S_2(\mathbf{Z}_{i2}; \gamma_2) \sigma_2^2 + 2\sigma_2^4 , \\ \text{var}(B_i \mid A_i^+ = 1, \mathbf{Z}_i) &= \mathbb{E}(L_{i1}^2 L_{i2}^2 \mid A_i^+ = 1, \mathbf{Z}_i; \mathbf{\Gamma}) - \mathbb{E}^2(L_{i1} L_{i2} \mid A_i^+ = 1, \mathbf{Z}_i; \mathbf{\Gamma}) \\ &= S_1^2 \sigma_2^2 + S_2^2 \sigma_1^2 + \sigma_1^2 \sigma_2^2 + 2\rho_1 S_1 S_2 \sigma_1 \sigma_2 + \rho_1^2 \sigma_1^2 \sigma_2^2 .\end{aligned}$$

We have then estimated $\hat{\gamma}_0$ and $\hat{\mathbf{\Gamma}}$ by solving $U_{11}(L_1, L_2 \mid A^+ = 1, \mathbf{Z}; \mathbf{\Gamma}) = 0$ and $U_{12}(A^+ \mid \mathbf{Z}_0; \gamma_0) = 0$. For estimation of $\mathbf{\Gamma}^\dagger$ and $\gamma_0^\dagger$, we write the estimating function in the form of

$$U_{21}(L_1, L_2 \mid A^+ = 1; \mathbf{\Gamma}^\dagger) = \sum_{i=1}^n U_{i21}(L_{i1}, L_{i2} \mid A_i^+ = 1; \mathbf{\Gamma}^\dagger) , \quad (\text{A.17})$$

$$U_{22}(A^+; \gamma_0^\dagger) = \sum_{i=1}^n U_{i22}(A_i^+; \gamma_0^\dagger) , \quad (\text{A.18})$$

where

$$U_{i21}(L_{i1}, L_{i2} \mid A_i^+ = 1; \mathbf{\Gamma}^\dagger) = A_i^+ D_i^{\dagger'} \Sigma_i^{\dagger-1} (W_i - \Omega_i^\dagger)$$

and $U_{i22}(A_i^+; \gamma_0^\dagger) = A_i^+ - \pi_{i2}(\gamma_0^\dagger)$, $\pi_2(\gamma_0^\dagger) = \mathbb{E}_{\mathbf{Z}_0}\{\pi_1(\mathbf{Z}_0; \gamma_0)\}$. Estimates of $\mathbf{\Gamma}^\dagger$ and $\gamma_0^\dagger$ can be obtained by solving the GEE2 equations with $D_i^\dagger$, $\Sigma_i^\dagger$ and $\Omega_i^\dagger$ defined analogously to D_i , Σ_i and Ω_i but without conditioning on $\mathbf{Z}$. This leads to

$$U_i(\boldsymbol{\Theta}) = \begin{pmatrix} U_{i0}(Y_i | \mathbf{X}_i; \boldsymbol{\varphi}, \mathbf{\Gamma}, \mathbf{\Gamma}^\dagger, \gamma_0, \gamma_0^\dagger) \\ U_{i11}(L_{i1}, L_{i2} | A_i^+ = 1, \mathbf{Z}_i; \mathbf{\Gamma}) \\ U_{i12}(A_i^+ | \mathbf{Z}_{i0}; \gamma_0) \\ U_{i21}(L_{i1}, L_{i2} | A_i^+ = 1; \mathbf{\Gamma}^\dagger) \\ U_{i22}(A_i^+; \gamma_0^\dagger) \end{pmatrix},$$

where $\boldsymbol{\Theta} = (\boldsymbol{\varphi}', \mathbf{\Gamma}', \mathbf{\Gamma}^{\dagger'}, \gamma_0', \gamma_0^{\dagger'})'$. We solve for $\mathbf{\Gamma}, \mathbf{\Gamma}^\dagger, \gamma_0$ and $\gamma_0^\dagger$ first and substitute these estimates into (A.14) to get an estimate for $\boldsymbol{\varphi}$, denoted by $\hat{\boldsymbol{\varphi}}$. The entire set of estimates is denoted by $\hat{\boldsymbol{\Theta}}$.

A.3 Covariance of Causal Effects of Multiple Exposures

Regression Adjustment

Here we derive the asymptotic covariance of the estimator obtained from both regression adjustment and inverse density weighting approaches. We use the semi-continuous exposures as an illustrative example, and the asymptotic covariance of the estimator obtained from binary exposures and continuous exposures can be derived in a similar fashion.

For regression adjustment, we have $\mathbb{E}(Y | \mathbf{X}, \mathbf{S}) = \mathbf{X}'\boldsymbol{\lambda} + \mathbf{S}'\boldsymbol{\eta}$ as shown in (2.4) where $\mathbf{X} = (1, A^+, A^+L_1, A^+L_2, A^+L_1L_2)'$ for semi-continuous exposures. We rewrite (2.4) as $\mathbb{E}\{Y | H\} = H'\boldsymbol{\beta}$, where $H = (1, \mathbf{X}', \mathbf{S}')'$. The ordinary least squares estimator of $\boldsymbol{\beta}$ is expressed as

$$\hat{\boldsymbol{\beta}} = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{Y},$$

where $\mathbf{Y}$ denotes the vector of the response for all n subjects in the sample, $\mathbf{H}$ denotes the matrix of exposures and generalized propensity scores for all n subjects in the sample.

Using estimating equation theory, the estimator $\boldsymbol{\beta}$ has covariance with the form of

$$\frac{1}{n}A^{-1}(\boldsymbol{\beta})B(\boldsymbol{\beta})\{A^{-1}(\boldsymbol{\beta})\}', \quad (\text{A.19})$$

where

$$A(\boldsymbol{\beta}) = \mathbb{E}\left\{-\frac{\partial H(Y - H'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\} = \mathbb{E}(HH') , \quad (\text{A.20})$$

and (A.20) can be estimated by

$$\hat{A}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n H_i H_i' .$$

We also have

$$\begin{aligned} B(\boldsymbol{\beta}) &= \text{var}\{H(Y - H'\boldsymbol{\beta})\} \\ &= \mathbb{E}\{\text{var}\{H(Y - H'\boldsymbol{\beta}) \mid H\}\} + \text{var}\{\mathbb{E}\{H(Y - H'\boldsymbol{\beta}) \mid H\}\} \\ &= \mathbb{E}\{\text{var}\{H(Y - H'\boldsymbol{\beta}) \mid H\}\} \\ &= \mathbb{E}\{\text{var}(R \mid H)HH'\} , \end{aligned}$$

where R might have a variance that varies with H . In this case, we estimate $\text{var}(R_i \mid H_i)$ by $\widehat{\text{var}}(R_i \mid H_i) = (Y_i - H_i'\hat{\boldsymbol{\beta}})^2$. Then the matrix $B(\boldsymbol{\beta})$ can be estimated by

$$\hat{B}(\hat{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^n H_i H_i' \widehat{\text{var}}(R_i \mid H_i) = \frac{1}{n} \sum_{i=1}^n H_i H_i' (Y_i - H_i'\hat{\boldsymbol{\beta}})^2 .$$

Thus, the robust sandwich variance estimator for linear regression is then given by

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \frac{1}{n} \mathbb{E}^{-1}(HH') \mathbb{E}\{\text{var}(R \mid H)HH'\} \{\mathbb{E}^{-1}(HH')\}' ,$$

which is estimated empirically by

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^n H_i H_i'\right)^{-1} \sum_{i=1}^n H_i H_i' (Y_i - H_i'\hat{\boldsymbol{\beta}})^2 \left\{\left(\sum_{i=1}^n H_i H_i'\right)^{-1}\right\}' .$$

Inverse Density Weighting

Now we derive the asymptotic covariance for the estimator obtained from the inverse density weighting approach with stabilized weights. Suppose $\hat{\boldsymbol{\varphi}}$ is the solution to $\mathbb{E}\{U_0(Y \mid \mathbf{X}; \boldsymbol{\varphi}, \boldsymbol{\Gamma}, \boldsymbol{\Gamma}^\dagger, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}_0^\dagger)\} = 0$, and $\boldsymbol{\varphi}$ along with other parameters can be estimated through second-order generalized estimating equations (Prentice and Zhao, 1991). We denote the

estimates of parameters $\Theta = (\varphi', \Gamma', \Gamma^{\dagger'}, \gamma'_0, \gamma_0^{\dagger'})'$ as $\hat{\Theta}$, and the stack of the estimating equations as $U(\Theta)$.

By Taylor series expansion,

$$U(\hat{\Theta}) = U(\Theta) + \frac{\partial U(\Theta)}{\partial \Theta'}(\hat{\Theta} - \Theta) + o_p(n^{-1/2}) .$$

Thus we have

$$\sqrt{n}(\hat{\Theta} - \Theta) = \left\{ -\frac{1}{n} \frac{\partial U(\Theta)}{\partial \Theta'} \right\}^{-1} \left\{ \frac{1}{\sqrt{n}} U(\Theta) \right\} + o_p(1) ,$$

where

$$\frac{\partial U(\Theta)}{\partial \Theta'} = \sum_{i=1}^n \frac{\partial U_i(\Theta)}{\partial \Theta'} ,$$

and

$$\frac{\partial U_i(\Theta)}{\partial \Theta'} = \begin{pmatrix} \frac{\partial U_{i0}}{\partial \varphi'} & \frac{\partial U_{i0}}{\partial \Gamma'} & \frac{\partial U_{i0}}{\partial \Gamma^{\dagger'}} & \frac{\partial U_{i0}}{\partial \gamma'_0} & \frac{\partial U_{i0}}{\partial \gamma_0^{\dagger'}} \\ 0 & \frac{\partial U_{i11}}{\partial \Gamma'} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\partial U_{i12}}{\partial \gamma'_0} & 0 \\ 0 & 0 & \frac{\partial U_{i21}}{\partial \Gamma^{\dagger'}} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\partial U_{i22}}{\partial \gamma_0^{\dagger'}} \end{pmatrix} .$$

As $n \rightarrow \infty$,

$$\sqrt{n}(\hat{\Theta} - \Theta) \xrightarrow{D} MNV(\mathbf{0}, \Sigma(\Theta)) ,$$

where

$$\Sigma(\Theta) = \mathcal{I}^{-1}(\Theta) \mathcal{C}(\Theta) [\mathcal{I}^{-1}(\Theta)]' ,$$

$$\mathcal{I}(\Theta) = \mathbb{E} \left\{ -\frac{\partial U_i(\Theta)}{\partial \Theta'} \right\} ,$$

and

$$\mathcal{C}(\Theta) = \mathbb{E} \{ U_i(\Theta) U_i'(\Theta) \} .$$

In practice, we use sample averages to estimate $\mathcal{I}(\Theta)$ and $\mathcal{C}(\Theta)$, and evaluate the resulting expressions at the estimates to obtain

$$\Sigma(\hat{\Theta}) = \mathcal{I}^{-1}(\hat{\Theta}) \mathcal{C}(\hat{\Theta}) \left\{ \mathcal{I}^{-1}(\hat{\Theta}) \right\}' , \quad (\text{A.21})$$

where

$$\mathcal{I}(\hat{\Theta}) = -\frac{1}{n} \sum_{i=1}^n \frac{\partial U_i(\Theta)}{\partial \Theta'} \bigg|_{\Theta=\hat{\Theta}},$$

and

$$\mathcal{C}(\hat{\Theta}) = \frac{1}{n} \sum_{i=1}^n U_i(\Theta) U_i'(\Theta) \bigg|_{\Theta=\hat{\Theta}}.$$

The upper left 5×5 matrix of (A.21) gives the estimated covariance matrix for $\hat{\varphi}$.

A.4 Supplementary Figures and Tables

Figures A.1 and A.2 present the relative biases of semi-continuous exposure effects estimated via regression approach and inverse density weighting approach under two model misspecifications as discussed in Section 2.4.3. Table A.1 provides a summary of the patient characteristics collected for the Detroit cohort study of prenatal alcohol exposure (PAE) and children’s cognition discussed in Section 2.6. Table A.2 presents a summary of assessing the covariate balance in the Detroit cohort study measured through maximum absolute correlation and average absolute correlation for different analyses discussed in Section 2.6.

Table A.1: Summary of characteristics of subjects in Detroit cohort study of prenatal alcohol exposure (PAE) and children cognition.

Characteristic	(n=377)
Smoking during pregnancy (cigarettes/day) (mean (SD))	9.56 (11.44)
Socioeconomic status at birth/infancy (mean (SD))	19.87 (8.59)
Mother’s age at delivery (mean (SD))	26.61 (6.08)
Prenatal Beck Depression Inventory (mean (SD))	11.05 (7.61)
Biological mother’s verbal IQ (PPVT score) (mean (SD))	72.64 (12.70)
Biological mother’s marital status (%)	
Not married	336 (89.1)
Married	41 (10.9)
Prenatal cocaine exposure (days/month) (mean (SD))	1.07 (2.87)
Prenatal marijuana exposure (days/month) (mean (SD))	0.75 (2.43)

Characteristic	(n=377)
Prenatal opiate exposure (days/month) (mean (SD))	0.19 (1.11)
Number of prenatal visits (mean (SD))	5.27 (3.28)
Biological mother's education (year) (mean (SD))	11.69 (1.59)
Parity (mean (SD))	1.36 (1.51)
Gravidity (mean (SD))	3.58 (2.33)
Gestational age at screening (mean (SD))	23.42 (7.78)
Child's gender (%)	
Male	222 (58.9)
Female	155 (41.1)
HOME score at infancy (mean (SD))	31.11 (4.80)
Beck Depression Inventory at 6-month postpartum follow-up (mean (SD))	11.31 (7.86)
Beck Depression Inventory at 12-month postpartum follow-up (mean (SD))	10.43 (7.48)
Socioeconomic status at age 7-year follow-up (mean (SD))	25.72 (11.14)
Beck Depression Inventory at 7-year follow-up (mean (SD))	8.23 (7.21)
HOME score at 7-year follow-up (mean (SD))	33.68 (6.94)
Primary caregiver at age 7 (%)	
Biological mother	55 (14.6)
Not biological mother	322 (85.4)
Primary caregiver's verbal IQ (PPVT score) at 7-year follow-up (mean (SD))	73.79 (14.02)
Primary caregiver's marital status at 7-year follow-up (%)	
Not married	304 (80.6)
Married	73 (19.4)
Number of stressful events at 7-year follow-up (mean (SD))	9.72 (5.38)
Perceived life stress at 7-year follow-up (mean (SD))	34.46 (26.58)
Socioeconomic status at age 14-year follow-up (mean (SD))	29.45 (10.21)
Primary caregiver at age 14 (%)	
Biological mother	70 (18.6)
Not biological mother	307 (81.4)
Family income at 14-year follow-up (mean (SD))	34231.42 (27408.45)
Primary caregiver's marital status at 14-year follow-up (%)	
Not married	280 (74.3)
Married	97 (25.7)
Beck Depression Inventory at 14-year follow-up (mean (SD))	10.65 (11.61)

Characteristic	(n=377)
Number of stressful events at 14-year follow-up (mean (SD))	7.22 (4.71)
Perceived life stress at 14-year follow-up (mean (SD))	32.61 (23.17)

Table A.2: Summary of covariate balance in terms of maximum absolute correlation and average absolute correlation.

Analysis	Method	Maximum absolute correlation	Average absolute correlation
Subgroup analysis	Covariate adjustment	0.41	0.12
	GPS adjustment	0.03	0.01
Full sample analysis with conventional regression	Covariate adjustment	0.37	0.09
Full sample analysis with two-part model	GPS adjustment	0.04	0.01
	IDW	0.08	0.04
Full sample analysis with two-stage approach	Reg Z + Reg Z	0.08	0.02
	Reg GPS + Reg Z	0.08	0.02
	Reg GPS + Reg PS	0.04	0.03
	Reg GPS + IPW	0.05	0.02

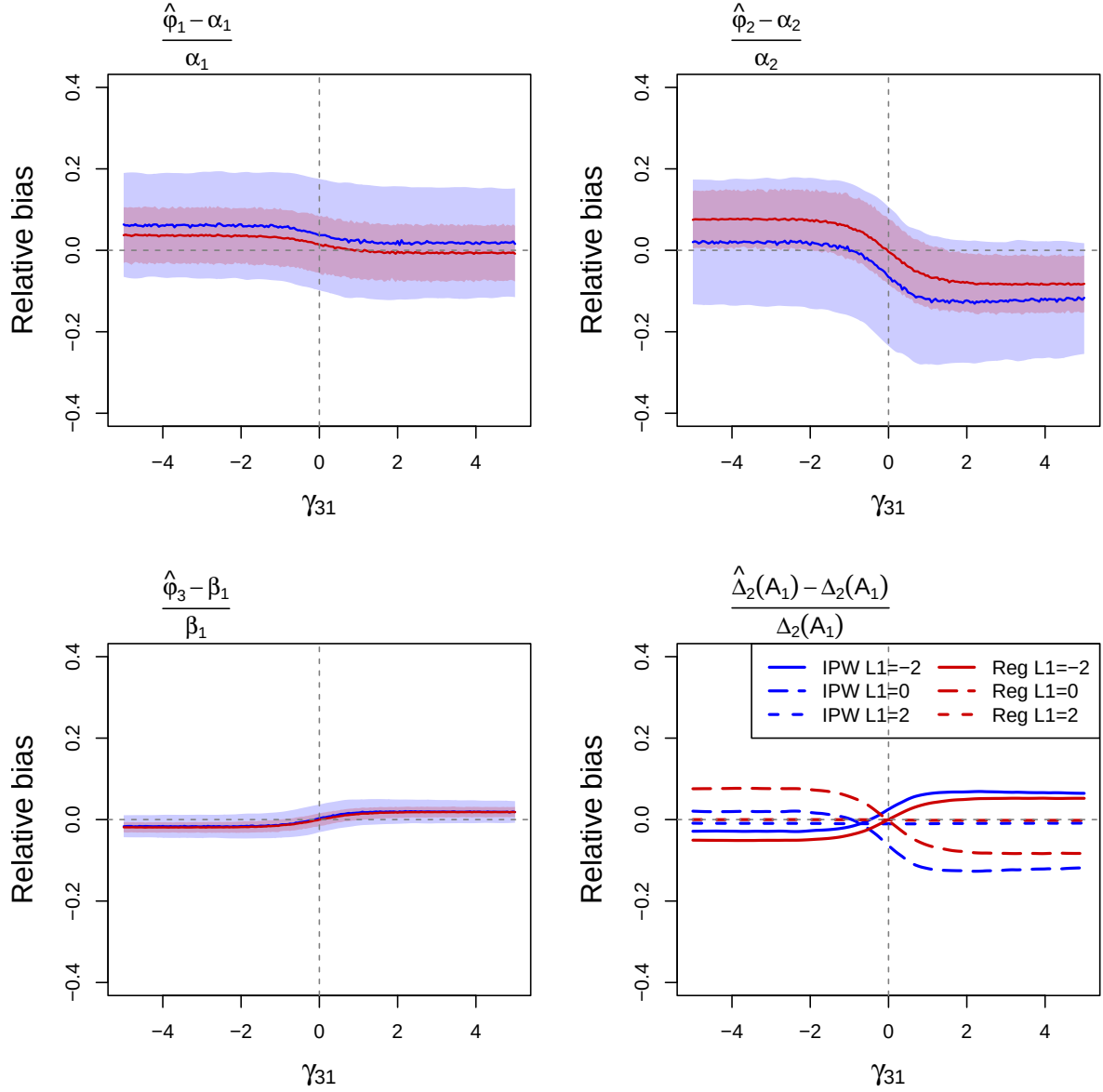


Figure A.1: Empirical relative bias with standard errors (shaded bands) of estimates of exposure effects from inadequate propensity score modeling assuming conditional independence between two semi-continuous exposures when the second-order confounder has the effect γ_{31} ; regression adjustment (Reg) (red) and inverse weighting (IPW) (blue) are considered; the marginal association between two semi-continuous exposures is $\mathbb{E}(\rho) = 0.5$.

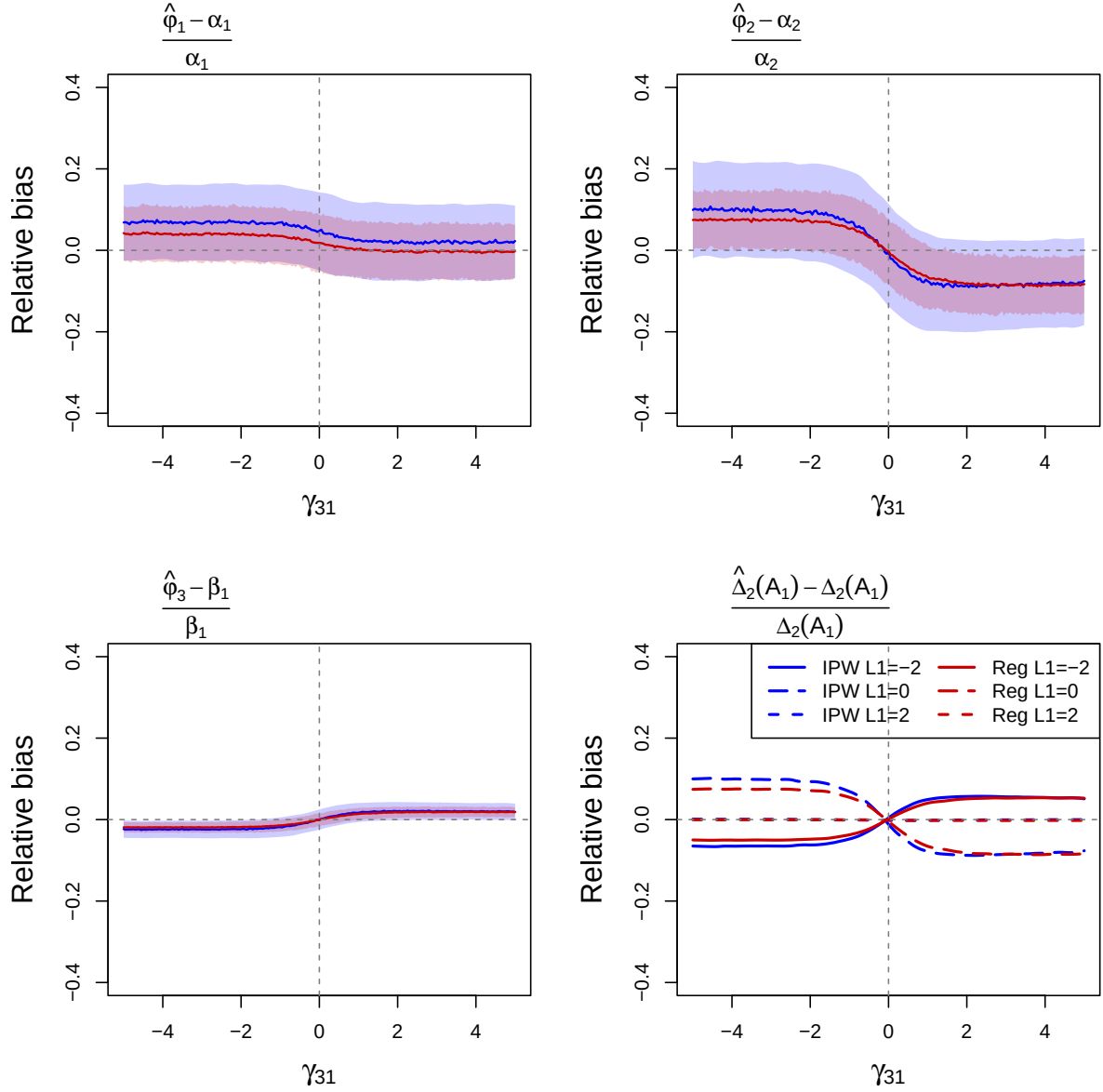


Figure A.2: Empirical relative bias with standard errors (shaded bands) of estimates of exposure effects from inadequate propensity score modeling due to the omission of a second-order confounder with effect γ_{31} ; regression adjustment (Reg) (red) and inverse weighting (IPW) (blue) are considered; the marginal association between two semi-continuous exposures is $\mathbb{E}(\rho) = 0.5$.

Appendix B

Appendix to Chapter 3

B.1 Derivations of Asymptotic Variance of $\hat{\beta}_1$ in a Simplified Setting

The simplified setting is one where $\mathbf{Z}_{ik} = (1, X_i)'$ for $k = 1, \dots, K$ (i.e. there is no covariate adjustment for the response model) and the mean response is the same over repeated measurements so $\mu(\mathbf{Z}_{ik}; \boldsymbol{\beta}) = \mu(X_i; \boldsymbol{\beta})$. We therefore replace $\mathbf{Z}_{ik}$ and $\mathbf{Z}_i$ by X_i to obtain the following. First

$$D(X_i; \boldsymbol{\beta}) = \frac{\boldsymbol{\mu}(X_i; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = \mu(X_i; \boldsymbol{\beta}) \begin{pmatrix} 1 & X_i \\ \vdots & \vdots \\ 1 & X_i \end{pmatrix}$$

is a $K_i \times p$ matrix. Second, with the Poisson variance function and working independence assumption the $K_i \times K_i$ working covariance matrix has identical entries in the diagonal so that

$$\Sigma(X_i; \boldsymbol{\beta}) = \begin{pmatrix} \mu(X_i; \boldsymbol{\beta}) & & 0 \\ & \ddots & \\ 0 & & \mu(X_i; \boldsymbol{\beta}) \end{pmatrix}.$$

The generalized estimating function in (3.2) then simplifies to

$$\begin{aligned} \mathbf{U}_i(X_i; \boldsymbol{\beta}) &= \mu(X_i; \boldsymbol{\beta}) \begin{pmatrix} 1 & \cdots & 1 \\ X_i & \cdots & X_i \end{pmatrix} \begin{pmatrix} \mu^{-1}(X_i; \boldsymbol{\beta}) & & 0 \\ & \ddots & \\ 0 & & \mu^{-1}(X_i; \boldsymbol{\beta}) \end{pmatrix} \begin{pmatrix} Y_{i1} - \mu(X_i; \boldsymbol{\beta}) \\ \vdots \\ Y_{i,K_i} - \mu(X_i; \boldsymbol{\beta}) \end{pmatrix} \\ &= \begin{pmatrix} 1 & \cdots & 1 \\ X_i & \cdots & X_i \end{pmatrix} \begin{pmatrix} Y_{i1} - \mu(X_i; \boldsymbol{\beta}) \\ \vdots \\ Y_{i,K_i} - \mu(X_i; \boldsymbol{\beta}) \end{pmatrix} = \sum_{j=1}^{K_i} \begin{pmatrix} U_{ij1}(\boldsymbol{\beta}) \\ U_{ij2}(\boldsymbol{\beta}) \end{pmatrix}, \end{aligned}$$

where $U_{ij1}(\boldsymbol{\beta}) = Y_{ij} - \mu(X_i; \boldsymbol{\beta})$ and $U_{ij2}(\boldsymbol{\beta}) = X_i(Y_{ij} - \mu(X_i; \boldsymbol{\beta}))$.

THE $\mathcal{A}(\boldsymbol{\beta})$ MATRIX

The matrix $\mathcal{A}(\boldsymbol{\beta})$ is given by

$$\begin{aligned} \mathcal{A}(\boldsymbol{\beta}) &= \mathbb{E} \{ D'(X_i; \boldsymbol{\beta}) \Sigma^{-1}(X_i; \boldsymbol{\beta}) D(X_i; \boldsymbol{\beta}) \} \\ &= \mathbb{E} \left\{ \mu(X_i; \boldsymbol{\beta}) \begin{pmatrix} 1 & \cdots & 1 \\ X_i & \cdots & X_i \end{pmatrix} \begin{pmatrix} \mu^{-1}(X_i; \boldsymbol{\beta}) & & 0 \\ & \ddots & \\ 0 & & \mu^{-1}(X_i; \boldsymbol{\beta}) \end{pmatrix} \mu(X_i; \boldsymbol{\beta}) \begin{pmatrix} 1 & X_i \\ \vdots & \vdots \\ 1 & X_i \end{pmatrix} \right\} \\ &= \mathbb{E} \left\{ \begin{pmatrix} 1 & \cdots & 1 \\ X_i & \cdots & X_i \end{pmatrix} \mu(X_i; \boldsymbol{\beta}) \begin{pmatrix} 1 & X_i \\ \vdots & \vdots \\ 1 & X_i \end{pmatrix} \right\} = \mathbb{E} \left\{ \begin{pmatrix} K_i \mu(X_i; \boldsymbol{\beta}) & K_i X_i \mu(X_i; \boldsymbol{\beta}) \\ K_i X_i \mu(X_i; \boldsymbol{\beta}) & K_i X_i^2 \mu(X_i; \boldsymbol{\beta}) \end{pmatrix} \right\}. \end{aligned}$$

Let $\kappa_x = \mathbb{E}(K_i \mid X_i = x)$ and $\nu_x = \text{var}(K_i \mid X_i = x)$ be the mean and variance of the number donations per subject in group x , $x = 0, 1$. Let $\mu_x = \mu(X_i = x; \boldsymbol{\beta})$ and let $V_x = \text{var}(Y_{ij} \mid X_i = x)$ be the true variance of the marginal responses for each subject in group x with $V_x = \mu_x(1 - \mu_x)$, $x = 0, 1$. Furthermore if we assume an exchangeable pairwise correlation for responses within individuals (i.e. $\text{corr}(Y_{ij}, Y_{ik} \mid X_i = x) = \rho$), then $\text{cov}(Y_{ij}, Y_{ik} \mid X_i = x) = \rho V_x$. Taking the expectation with respect to K_i and X_i we obtain

$$\mathcal{A}(\boldsymbol{\beta}) = \begin{pmatrix} \sum_{x=0}^1 \kappa_x \mu_x P(X_i = x) & \kappa_1 \mu_1 P(X_i = 1) \\ \kappa_1 \mu_1 P(X_i = 1) & \kappa_1 \mu_1 P(X_i = 1) \end{pmatrix},$$

with inverse

$$\mathcal{A}^{-1}(\boldsymbol{\beta}) = \begin{pmatrix} \frac{1}{\kappa_0 \mu_0 P(X_i = 0)} & -\frac{1}{\kappa_0 \mu_0 P(X_i = 0)} \\ -\frac{1}{\kappa_0 \mu_0 P(X_i = 0)} & \frac{1}{\kappa_0 \mu_0 P(X_i = 0)} + \frac{1}{\kappa_1 \mu_1 P(X_i = 1)} \end{pmatrix}.$$

THE $\mathcal{B}(\boldsymbol{\beta})$ MATRIX

The matrix $\mathcal{B}(\boldsymbol{\beta}) = \mathbb{E}\{\mathbf{U}_i(X_i; \boldsymbol{\beta}) \mathbf{U}_i'(X_i; \boldsymbol{\beta})\}$ can be written as

$$\mathcal{B}(\boldsymbol{\beta}) = \mathbb{E} \left\{ \begin{array}{cc} \sum_{j=1}^{K_i} \sum_{k=1}^{K_i} U_{ij1}(\boldsymbol{\beta}) U_{ik1}(\boldsymbol{\beta}) & \sum_{j=1}^{K_i} \sum_{k=1}^{K_i} U_{ij1}(\boldsymbol{\beta}) U_{ik2}(\boldsymbol{\beta}) \\ \sum_{j=1}^{K_i} \sum_{k=1}^{K_i} U_{ij1}(\boldsymbol{\beta}) U_{ik2}(\boldsymbol{\beta}) & \sum_{j=1}^{K_i} \sum_{k=1}^{K_i} U_{ij2}(\boldsymbol{\beta}) U_{ik2}(\boldsymbol{\beta}) \end{array} \right\}. \quad (\text{B.1})$$

For the (1, 1) entry of (B.1) note that

$$\begin{aligned} \mathbb{E} \left\{ \sum_{j=1}^{K_i} \sum_{k=1}^{K_i} U_{ij1}(\boldsymbol{\beta}) U_{ik1}(\boldsymbol{\beta}) \mid X_i, K_i \right\} &= \mathbb{E} \left\{ \sum_{j=1}^{K_i} U_{ij1}^2(\boldsymbol{\beta}) + \sum_{\substack{j=1 \\ j \neq k}}^{K_i} \sum_{k=1}^{K_i} U_{ij1}(\boldsymbol{\beta}) U_{ik1}(\boldsymbol{\beta}) \mid X_i, K_i \right\} \\ &= \left\{ K_i \text{var}(Y_{ij} \mid X_i) + \sum_{\substack{j=1 \\ j \neq k}}^{K_i} \sum_{k=1}^{K_i} \text{cov}(Y_{ij}, Y_{ik} \mid X_i) \right\} \\ &= K_i \text{var}(Y_{ij} \mid X_i) + K_i (K_i - 1) \text{cov}(Y_{ij}, Y_{ik} \mid X_i). \end{aligned}$$

This has expectation

$$\begin{aligned} &\sum_{x=0}^1 \left[\mathbb{E}(K_i \mid X_i = x) \text{var}(Y_{ij} \mid X_i = x) P(X_i = x) + \mathbb{E}\{K_i(K_i - 1) \mid X_i = x\} \text{cov}(Y_{ij}, Y_{ik} \mid X_i = x) P(X_i = x) \right] \\ &= \sum_{x=0}^1 \left\{ \kappa_x \text{var}(Y_{ij} \mid X_i = x) + (\nu_x + \kappa_x^2 - \kappa_x) \text{cov}(Y_{ij}, Y_{ik} \mid X_i = x) \right\} P(X_i = x) \\ &= \sum_{x=0}^1 \left\{ \kappa_x V_x + (\nu_x + \kappa_x^2 - \kappa_x) \rho V_x \right\} P(X_i = x). \end{aligned}$$

For the (2, 2) entry of (B.1) we have

$$\begin{aligned}
& \mathbb{E} \left\{ \sum_{j=1}^{K_i} \sum_{k=1}^{K_i} U_{ij1}(\boldsymbol{\beta}) U_{ik2}(\boldsymbol{\beta}) \mid X_i, K_i \right\} \\
&= \mathbb{E} \left\{ \sum_{j=1}^{K_i} \sum_{k=1}^{K_i} (Y_{ij} - \mu(X_i; \boldsymbol{\beta})) X_i (Y_{ik} - \mu(X_i; \boldsymbol{\beta})) \mid X_i, K_i \right\} \\
&= \mathbb{E} \left\{ \sum_{j=1}^{K_i} X_i U_{ij1}^2 + \sum_{\substack{j=1 \\ j \neq k}}^{K_i} \sum_{k=1}^{K_i} X_i (Y_{ij} - \mu(X_i; \boldsymbol{\beta})) (Y_{ik} - \mu(X_i; \boldsymbol{\beta})) \mid X_i, K_i \right\} \\
&= \{K_i \text{var}(Y_{ij} \mid X_i = 1) + K_i(K_i - 1) \text{cov}(Y_{ij}, Y_{ik} \mid X_i = 1)\} I(X_i = 1).
\end{aligned}$$

Taking the expectation with respect to K_i and X_i gives

$$\begin{aligned}
& \{\kappa_1 \text{var}(Y_{ij} \mid X_i = 1) + (\nu_1 + \kappa_1^2 - \kappa_1) \text{cov}(Y_{ij}, Y_{ik} \mid X_i = 1)\} P(X_i = 1) = \\
& \{\kappa_1 V_1 + (\nu_1 + \kappa_1^2 - \kappa_1) \rho V_1\} P(X_i = 1).
\end{aligned}$$

For the (1, 2) and (2, 1) entries given X_i and K_i we get

$$\{K_i \text{var}(Y_{ij} \mid X_i = 1) + K_i(K_i - 1) \text{cov}(Y_{ij}, Y_{ik} \mid X_i = 1)\} I(X_i = 1),$$

which is the same as the (2, 2) entry so we have

$$\mathcal{B}(\boldsymbol{\beta}) = \begin{pmatrix} \sum_{x=0}^1 \{\kappa_x V_x + (\nu_x + \kappa_x^2 - \kappa_x) \rho V_x\} P(X_i = x) & \{\kappa_1 V_1 + (\nu_1 + \kappa_1^2 - \kappa_1) \rho V_1\} P(X_i = 1) \\ \{\kappa_1 V_1 + (\nu_1 + \kappa_1^2 - \kappa_1) \rho V_1\} P(X_i = 1) & \{\kappa_1 V_1 + (\nu_1 + \kappa_1^2 - \kappa_1) \rho V_1\} P(X_i = 1) \end{pmatrix}.$$

Thus $\mathcal{A}^{-1}(\boldsymbol{\beta}) \mathcal{B}(\boldsymbol{\beta}) \{\mathcal{A}^{-1}(\boldsymbol{\beta})\}'$ is

$$\begin{pmatrix} \frac{V_0[\kappa_0 + (\nu_0 + \kappa_0^2 - \kappa_0)\rho]}{\kappa_0^2 \mu_0^2 P(X_i = 0)} & -\frac{V_0[\kappa_0 + (\nu_0 + \kappa_0^2 - \kappa_0)\rho]}{\kappa_0^2 \mu_0^2 P(X_i = 0)} \\ -\frac{V_0[\kappa_0 + (\nu_0 + \kappa_0^2 - \kappa_0)\rho]}{\kappa_0^2 \mu_0^2 P(X_i = 0)} & \sum_{x=0}^1 \frac{V_x[\kappa_x + (\nu_x + \kappa_x^2 - \kappa_x)\rho]}{\kappa_x^2 \mu_x^2 P(X_i = x)} \end{pmatrix}, \quad (\text{B.2})$$

and the asymptotic variance of $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ is the (2, 2) entry of (B.2), i.e.,

$$\Omega_{\beta_1\beta_1} = \frac{V_0}{\kappa_0\mu_0^2P(X_i=0)} \left(1 + \frac{(\nu_0 + \kappa_0^2 - \kappa_0)\rho}{\kappa_0}\right) + \frac{V_1}{\kappa_1\mu_1^2P(X_i=1)} \left(1 + \frac{(\nu_1 + \kappa_1^2 - \kappa_1)\rho}{\kappa_1}\right).$$

B.2 Conditional Mean and Variance of $K_F \mid K_A \geq 1$

Let K_A denote the number of transfusions from a random subject during the accrual period of length τ_A , and K_F be the number of transfusions from the subject in the follow-up window of length τ_F . To be recruited we require $K_A \geq 1$ and we are then interested in the number of transfusions K_F following recruitment over the follow-up period of duration τ_F . Suppose $\mathbf{U} > 0$ is a random effect with a gamma distribution with $\mathbb{E}(\mathbf{U}) = 1$ and $\text{var}(\mathbf{U}) = \phi$ so that

$$dG(u; \phi) = \frac{u^{\phi^{-1}-1}e^{-u/\phi}}{\Gamma(\phi^{-1})\phi^{\phi^{-1}}}du. \quad (\text{B.3})$$

With the conditional Poisson models,

$$K_A \mid u \sim \text{Poisson}(u\lambda\tau_A) \text{ and } K_F \mid u \sim \text{Poisson}(u\lambda\tau_F),$$

then

$$dG(u \mid K_A \geq 1; \lambda, \phi) = \frac{P(K_A \geq 1 \mid u; \lambda)dG(u; \phi)}{\int_0^\infty P(K_A \geq 1 \mid u; \lambda)dG(u; \phi)} = \frac{(1 - e^{-u\lambda\tau_A})dG(u; \phi)}{\int_0^\infty (1 - e^{-u\lambda\tau_A})dG(u; \phi)}. \quad (\text{B.4})$$

To evaluate the conditional density of $\mathbf{U}$ the denominator of (B.4) is

$$\begin{aligned} \int_0^\infty (1 - e^{-u\lambda\tau_A})dG(u; \phi) &= 1 - \int_0^\infty e^{-u\lambda\tau_A} \frac{u^{\phi^{-1}-1}e^{-u/\phi}}{\Gamma(\phi^{-1})\phi^{\phi^{-1}}}du = 1 - \int_0^\infty \frac{u^{\phi^{-1}-1}e^{-u(1/\phi + \lambda\tau_A)}}{\Gamma(\phi^{-1})\phi^{\phi^{-1}}}du \\ &= 1 - \frac{\Gamma(\phi^{-1})}{\Gamma(\phi^{-1})\phi^{\phi^{-1}}} \left(\frac{\phi}{1 + \lambda\tau_A\phi} \right)^{\phi^{-1}} = 1 - \left(\frac{1}{1 + \lambda\tau_A\phi} \right)^{\phi^{-1}}. \end{aligned}$$

Alternatively since

$$P(K = k) = \frac{\Gamma\left(\frac{1}{\phi} + k\right)}{\Gamma\left(\frac{1}{\phi}\right)k!} \left(\frac{\lambda\tau\phi}{1 + \lambda\tau\phi} \right)^k \left(\frac{1}{1 + \lambda\tau\phi} \right)^{\phi^{-1}}$$

is the negative binomial probability mass function with mean $\mathbb{E}(K) = \lambda\tau$ and variance $\text{var}(K) = \lambda\tau(1 + \lambda\tau\phi)$, we can see that the denominator of (B.4) is

$$P(K_A \geq 1) = 1 - P(K_A = 0) = 1 - \left(\frac{1}{1 + \lambda\tau_A\phi} \right)^{\phi^{-1}}.$$

Then (B.4) becomes

$$\begin{aligned} dG(u \mid K_A \geq 1; \lambda, \phi) &= \frac{\frac{u^{\phi^{-1}-1} e^{-u/\phi}}{\Gamma(\phi^{-1})\phi^{\phi^{-1}}} du - \frac{u^{\phi^{-1}-1} e^{-u/\frac{\phi}{1+\lambda\tau_A\phi}}}{\Gamma(\phi^{-1})\phi^{\phi^{-1}}} du}{1 - \left(\frac{1}{1 + \lambda\tau_A\phi} \right)^{\phi^{-1}}} \\ &= \frac{(1 + \lambda\tau_A\phi)^{\phi^{-1}}}{(1 + \lambda\tau_A\phi)^{\phi^{-1}} - 1} \left\{ \frac{u^{\phi^{-1}-1} e^{-u/\phi}}{\Gamma(\phi^{-1})\phi^{\phi^{-1}}} \right\} du - \\ &\quad \frac{1}{(1 + \lambda\tau_A\phi)^{\phi^{-1}} - 1} \left\{ \frac{u^{\phi^{-1}-1} e^{-u/\frac{\phi}{1+\lambda\tau_A\phi}}}{\Gamma(\phi^{-1})\left(\frac{\phi}{1+\lambda\tau_A\phi}\right)^{\phi^{-1}}} \right\} du, \end{aligned}$$

which is a combination of two weighted gamma distributions, i.e.,

$$\text{Gamma}(\phi^{-1}, \phi) \text{ and } \text{Gamma}\left(\phi^{-1}, \frac{\phi}{1 + \lambda\tau_A\phi}\right).$$

Therefore

$$\begin{aligned} \mathbb{E}(u \mid K_A \geq 1) &= \frac{(1 + \lambda\tau_A\phi)^{\phi^{-1}}}{(1 + \lambda\tau_A\phi)^{\phi^{-1}} - 1} \phi^{-1} \phi - \frac{1}{(1 + \lambda\tau_A\phi)^{\phi^{-1}} - 1} \left(\phi^{-1} \frac{\phi}{1 + \lambda\tau_A\phi} \right) \\ &= \frac{1}{(1 + \lambda\tau_A\phi)^{\phi^{-1}} - 1} \frac{(1 + \lambda\tau_A\phi)^{\phi^{-1}+1} - 1}{1 + \lambda\tau_A\phi}, \end{aligned} \tag{B.5}$$

and

$$\begin{aligned}
\mathbb{E}(u^2 \mid K_A \geq 1) &= \frac{(1 + \lambda\tau_A\phi)^{\phi^{-1}}}{(1 + \lambda\tau_A\phi)^{\phi^{-1}} - 1} \left\{ \phi^{-1}\phi^2 + (\phi^{-1}\phi)^2 \right\} - \\
&\quad \frac{1}{(1 + \lambda\tau_A\phi)^{\phi^{-1}} - 1} \left\{ \phi^{-1} \left(\frac{\phi}{1 + \lambda\tau_A\phi} \right)^2 + \left(\phi^{-1} \frac{\phi}{1 + \lambda\tau_A\phi} \right)^2 \right\} \\
&= \frac{\phi + 1}{(1 + \lambda\tau_A\phi)^{\phi^{-1}} - 1} \left\{ (1 + \lambda\tau_A\phi)^{\phi^{-1}} - \frac{1}{(1 + \lambda\tau_A\phi)^2} \right\}. \tag{B.6}
\end{aligned}$$

The mean of K_F given $K_A \geq 1$ is

$$\mathbb{E}(K_F \mid K_A \geq 1) = \mathbb{E}_U \{ \mathbb{E}(K_F \mid \mathbf{U}, K_A \geq 1) \mid K_A \geq 1 \} = \lambda\tau_F \mathbb{E}(u \mid K_A \geq 1), \tag{B.7}$$

and the variance of K_F given $K_A \geq 1$ is

$$\begin{aligned}
\text{var}(K_F \mid K_A \geq 1) &= \text{var} \{ \mathbb{E}(K_F \mid u, K_A \geq 1) \mid K_A \geq 1 \} + \mathbb{E} \{ \text{var}(K_F \mid u, K_A \geq 1) \mid K_A \geq 1 \} \\
&= \text{var}(u\lambda\tau_F \mid K_A \geq 1) + \mathbb{E}(u\lambda\tau_F \mid K_A \geq 1) \\
&= (\lambda\tau_F)^2 \text{var}(u \mid K_A \geq 1) + \lambda\tau_F \mathbb{E}(u \mid K_A \geq 1) \\
&= (\lambda\tau_F)^2 \{ \mathbb{E}(u^2 \mid K_A \geq 1) - \mathbb{E}^2(u \mid K_A \geq 1) \} + \lambda\tau_F \mathbb{E}(u \mid K_A \geq 1). \tag{B.8}
\end{aligned}$$

Plugging (B.5) and (B.6) into (B.7) and (B.8) gives

$$\mathbb{E}(K_F \mid K_A \geq 1) = \lambda\tau_F \left\{ \frac{1 - (1 + \lambda\tau_A\phi)^{-\phi^{-1}-1}}{1 - (1 + \lambda\tau_A\phi)^{-\phi^{-1}}} \right\},$$

and

$$\text{var}(K_F \mid K_A \geq 1) = \mathbb{E}(K_F \mid K_A \geq 1) - \mathbb{E}^2(K_F \mid K_A \geq 1) + (\lambda\tau_F)^2 (1 + \phi) \left\{ \frac{1 - (1 + \lambda\tau_A\phi)^{-\phi^{-1}-2}}{1 - (1 + \lambda\tau_A\phi)^{-\phi^{-1}}} \right\}.$$

To estimate the parameters of the zero-truncated negative binomial model we consider

the likelihood function of the form

$$\begin{aligned}
L(\lambda, \phi) &\propto \prod_{i=1}^n P(K_{iF} = k_i \mid K_{iA} \geq 1; \lambda, \phi) \\
&= \prod_{i=1}^n \int_0^\infty P(K_{iF} = k_i \mid K_{iA} \geq 1, u; \lambda, \phi) P(u \mid K_{iA} \geq 1) du \\
&= \prod_{i=1}^n \int_0^\infty P(K_{iF} = k_i \mid u; \lambda, \phi) P(u \mid K_{iA} \geq 1) du
\end{aligned}$$

where $P(K_{iF} = k_i \mid u; \lambda, \phi) P(u \mid K_{iA} \geq 1) du$ can be written as

$$\begin{aligned}
&\int_0^\infty \frac{(u\lambda\tau_F)^{k_i} e^{-u\lambda\tau_F}}{k_i!} \left\{ \frac{(1 + \lambda\tau_A\phi)^{\phi^{-1}}}{(1 + \lambda\tau_A\phi)^{\phi^{-1}} - 1} \left\{ \frac{u^{\phi^{-1}-1} e^{-u/\phi}}{\Gamma(\phi^{-1})\phi^{\phi^{-1}}} \right\} - \right. \\
&\quad \left. \frac{1}{(1 + \lambda\tau_A\phi)^{\phi^{-1}} - 1} \left\{ \frac{u^{\phi^{-1}-1} e^{-u/(1+\lambda\tau_A\phi)}}{\Gamma(\phi^{-1})(\frac{\phi}{1+\lambda\tau_A\phi})^{\phi^{-1}}} \right\} \right\} du
\end{aligned}$$

which simplifies to

$$\begin{aligned}
&\frac{\Gamma(\phi^{-1} + k_i)}{\Gamma(\phi^{-1})k_i!} \frac{1}{(1 + \lambda\tau_A\phi)^{\phi^{-1}} - 1} \left\{ \left(\frac{\lambda\tau_F\phi}{1 + \lambda\tau_F\phi} \right)^{k_i} \left(\frac{1 + \lambda\tau_A\phi}{1 + \lambda\tau_F\phi} \right)^{\phi^{-1}} - \right. \\
&\quad \left. \left(\frac{\lambda\tau_F\phi}{1 + \lambda\tau_F\phi + \lambda\tau_A\phi} \right)^{k_i} \left(\frac{1 + \lambda\tau_A\phi}{1 + \lambda\tau_F\phi + \lambda\tau_A\phi} \right)^{\phi^{-1}} \right\}.
\end{aligned}$$

B.3 Conditional Mean and Variance of $K_F \mid L \leq \tau_A$

Let the recruitment time for an individual be L . We have $L \mid U \sim \exp(U\lambda)$ following the mixed Poisson model of the number of the donations. Marginalizing over U we have

$$f(l; \lambda, \phi) = \int_0^\infty P(l \mid u; \lambda) dG(u; \phi) = \int_0^\infty u \lambda \exp(-u\lambda l) \frac{u^{\phi^{-1}-1} e^{-u/\phi}}{\Gamma(\phi^{-1})\phi^{\phi^{-1}}} du = \lambda \left(\frac{1}{\lambda l \phi + 1} \right)^{\phi^{-1}+1},$$

and

$$P(L \leq \tau_A) = \int_0^{\tau_A} \lambda(\lambda l \phi + 1)^{-\phi^{-1}-1} dl = 1 - (\lambda \phi \tau_A + 1)^{-\phi^{-1}},$$

and

$$dG(u \mid l; \lambda, \phi)$$

$$= \frac{P(l \mid u; \lambda) dG(u; \phi)}{\int_0^\infty P(l \mid u; \lambda) dG(u; \phi)} = \frac{u \lambda \exp(-u \lambda l) \frac{u^{\phi^{-1}-1} e^{-u/\phi}}{\Gamma(\phi^{-1}) \phi^{\phi^{-1}}}}{\int_0^\infty u \lambda \exp(-u \lambda l) \frac{u^{\phi^{-1}-1} e^{-u/\phi}}{\Gamma(\phi^{-1}) \phi^{\phi^{-1}}} du} = \frac{u^{\phi^{-1}} \exp\left(-u / \frac{\phi}{\lambda l + 1}\right)}{\Gamma(\phi^{-1} + 1) \left(\frac{\phi}{\lambda l + 1}\right)^{\phi^{-1} + 1}},$$

which gives

$$U \mid L = l \sim \text{Gamma} \left(\text{shape} = \phi^{-1} + 1, \text{scale} = \frac{\phi}{\lambda l + 1} \right).$$

The mean of K_F given $L \leq \tau_A$ is

$$\mathbb{E}(K_F \mid L \leq \tau_A) = \mathbb{E}_U \{ \mathbb{E}(K_F \mid \mathbf{U}, L \leq \tau_A) \mid L \leq \tau_A \} = \lambda \tau_F \mathbb{E}(U \mid L \leq \tau_A),$$

and the variance of K_F given $L \leq \tau_A$ is

$$\begin{aligned} \text{var}(K_F \mid L \leq \tau_A) &= \text{var} \{ \mathbb{E}(K_F \mid U, L \leq \tau_A) \mid L \leq \tau_A \} + \mathbb{E} \{ \text{var}(K_F \mid U, L \leq \tau_A) \mid L \leq \tau_A \} \\ &= \text{var}(U \lambda \tau_F \mid L \leq \tau_A) + \mathbb{E}(U \lambda \tau_F \mid L \leq \tau_A) \\ &= (\lambda \tau_F)^2 \text{var}(U \mid L \leq \tau_A) + \lambda \tau_F \mathbb{E}(U \mid L \leq \tau_A) \\ &= (\lambda \tau_F)^2 \{ \mathbb{E}(U^2 \mid L \leq \tau_A) - \mathbb{E}^2(U \mid L \leq \tau_A) \} + \lambda \tau_F \mathbb{E}(U \mid L \leq \tau_A), \end{aligned}$$

where

$$\begin{aligned} \mathbb{E}(U \mid L \leq \tau_A) &= \frac{\int_0^{\tau_A} \int_0^\infty U f(U, l) du dl}{\int_0^{\tau_A} \int_0^\infty f(U, l) du dl} = \frac{\int_0^{\tau_A} \int_0^\infty U f(U \mid l) f(l) du dl}{\int_0^{\tau_A} \int_0^\infty f(U \mid l) f(l) du dl} \\ &= \frac{\int_0^{\tau_A} f(l) \{ \int_0^\infty U f(U \mid l) du \} dl}{\int_0^{\tau_A} f(l) \{ \int_0^\infty f(U \mid l) du \} dl} = \frac{\int_0^{\tau_A} \frac{1+\phi}{\lambda \phi l + 1} f(l) dl}{\int_0^{\tau_A} f(l) dl} = \frac{1 - (\lambda \phi \tau_A + 1)^{-\phi^{-1}-1}}{1 - (\lambda \phi \tau_A + 1)^{-\phi^{-1}}}, \end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}(U^2 \mid L \leq \tau_A) &= \frac{\int_0^{\tau_A} f(l) \left\{ \int_0^\infty U^2 f(U \mid l) du \right\} dl}{\int_0^{\tau_A} f(l) dl} \\
&= \frac{\int_0^{\tau_A} f(l) \left\{ (\phi^{-1} + 1) \left(\frac{\phi}{\lambda\phi l + 1} \right)^2 + \left(\frac{1+\phi}{\lambda\phi l + 1} \right)^2 \right\} dl}{\int_0^{\tau_A} f(l) dl} \\
&= \frac{(\phi + 1)(2\phi + 1) \int_0^{\tau_A} f(l) (\lambda\phi l + 1)^{-2} dl}{1 - (\lambda\phi\tau_A + 1)^{-\phi^{-1}}} = (1 + \phi) \frac{1 - (\lambda\phi\tau_A + 1)^{-\phi^{-1}-2}}{1 - (\lambda\phi\tau_A + 1)^{-\phi^{-1}}}.
\end{aligned}$$

Therefore we can show that

$$\mathbb{E}(K_F \mid L \leq \tau_A) = \lambda\tau_F \left\{ \frac{1 - (1 + \lambda\tau_A\phi)^{-\phi^{-1}-1}}{1 - (1 + \lambda\tau_A\phi)^{-\phi^{-1}}} \right\},$$

and

$$\text{var}(K_F \mid L \leq \tau_A) = \mathbb{E}(K_F \mid L \leq \tau_A) - \mathbb{E}^2(K_F \mid L \leq \tau_A) + (\lambda\tau_F)^2 (1 + \phi) \left\{ \frac{1 - (1 + \lambda\tau_A\phi)^{-\phi^{-1}-2}}{1 - (1 + \lambda\tau_A\phi)^{-\phi^{-1}}} \right\}.$$

Since we count the donation that led to the identification and recruitment of the individual we let $K = 1 + K_F$ and so $\kappa = 1 + \mathbb{E}(K_F \mid K_A \geq 1) = 1 + \mathbb{E}(K_F \mid L \leq \tau_A)$ and $\nu = \text{var}(K_F \mid K_A \geq 1) = \text{var}(K_F \mid L \leq \tau_A)$.

B.4 Matrices $\mathcal{A}(t_r)$, $\mathcal{B}(t_r)$ and $\mathcal{B}(t_r, t_s)$ in a Simplified Setting

Without the adjustment for covariates other than the assigned treatment, $\mathbf{Z}_{ik} = \mathbf{Z}_i = (1, X_i)'$ for $k = 1, \dots, K_i(t_r)$, and $\boldsymbol{\beta} = (\beta_1, \beta_0)'$. Hence the means of responses are the same over repeated measurements, i.e. $\mu(\mathbf{Z}_{ik}; \boldsymbol{\beta}) = \mu(X_i; \boldsymbol{\beta})$. If we let $\mu(X_i) = \mu(X_i; \boldsymbol{\beta})$ for simplification, and specifically with the log link $D_i(t_r; \boldsymbol{\beta})$ is a $K_i(t_r) \times 2$ matrix of the

form

$$D_i(t_r; \boldsymbol{\beta}) = \frac{\partial \boldsymbol{\mu}_i(t_r; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = \mu(X_i) \begin{pmatrix} X_i & 1 \\ \vdots & \vdots \\ X_i & 1 \end{pmatrix}. \quad (\text{B.9})$$

With the identity (Poisson) variance function and a working independence assumption, the $K_i(t_r) \times K_i(t_r)$ working covariance matrix has identical entries in the diagonal,

$$\Sigma_i^{-1}(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha}) = \Sigma_i^{-1}(t_r; \boldsymbol{\beta}) = \begin{pmatrix} \mu^{-1}(X_i) & & 0 \\ & \ddots & \\ 0 & & \mu^{-1}(X_i) \end{pmatrix}. \quad (\text{B.10})$$

The generalized estimating function in (3.18) then simplifies to

$$\begin{aligned} \boldsymbol{U}_i(t_r; \boldsymbol{\beta}) &= \Delta_i(t_r) \mu(X_i) \begin{pmatrix} X_i & \cdots & X_i \\ 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu^{-1}(X_i) & & 0 \\ & \ddots & \\ 0 & & \mu^{-1}(X_i) \end{pmatrix} \begin{pmatrix} Y_{i1} - \mu(X_i) \\ \vdots \\ Y_{i, K_i(t_r)} - \mu(X_i) \end{pmatrix} \\ &= \Delta_i(t_r) \begin{pmatrix} X_i & \cdots & X_i \\ 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} Y_{i1} - \mu(X_i) \\ \vdots \\ Y_{i, K_i(t_r)} - \mu(X_i) \end{pmatrix} \\ &= \Delta_i(t_r) \begin{pmatrix} \sum_{j=1}^{K_i(t_r)} X_i (Y_{ij} - \mu(X_i)) \\ \sum_{j=1}^{K_i(t_r)} (Y_{ij} - \mu(X_i)) \end{pmatrix} \\ &= \Delta_i(t_r) \sum_{j=1}^{K_i(t_r)} \begin{pmatrix} U_{ij1}(\boldsymbol{\beta}) \\ U_{ij2}(\boldsymbol{\beta}) \end{pmatrix}. \end{aligned}$$

Derivation of the $\mathcal{A}(t_r)$ Matrix

The $\mathcal{A}(t_r)$ is given more explicitly by

$$\begin{aligned}
\mathcal{A}(t_r) &= \mathbb{E} \left\{ \Delta_i(t_r) D_i'(t_r; \boldsymbol{\beta}) \Sigma_i^{-1}(t_r; \boldsymbol{\beta}, \boldsymbol{\alpha}) D_i(t_r; \boldsymbol{\beta}) \right\} \\
&= \mathbb{E} \left\{ \Delta_i(t_r) \mu(X_i) \begin{pmatrix} X_i & \cdots & X_i \\ 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu^{-1}(X_i) & & 0 \\ & \ddots & \\ 0 & & \mu^{-1}(X_i) \end{pmatrix} \mu(X_i) \begin{pmatrix} X_i & 1 \\ \vdots & \vdots \\ X_i & 1 \end{pmatrix} \right\} \\
&= \mathbb{E} \left\{ \Delta_i(t_r) \begin{pmatrix} X_i & \cdots & X_i \\ 1 & \cdots & 1 \end{pmatrix} \mu(X_i) \begin{pmatrix} X_i & 1 \\ \vdots & \vdots \\ X_i & 1 \end{pmatrix} \right\} \\
&= \mathbb{E} \left\{ \Delta_i(t_r) \begin{pmatrix} K_i(t_r) X_i \mu(X_i) & K_i(t_r) X_i \mu(X_i) \\ K_i(t_r) X_i \mu(X_i) & K_i(t_r) \mu(X_i) \end{pmatrix} \right\}.
\end{aligned}$$

Let $\kappa_x(t_r) = \mathbb{E}\{K_i(t_r) \mid X_i = x, \Delta_i(t_r) = 1\}$ be the expected number of measurements for a subject in group x for the r th interim analysis. Let $K_i(t_r, t_s)$ be the number of transfusions needed over time period $(t_r, t_s]$ for individual i , then $\kappa_x(t_r, t_s) = \mathbb{E}\{K_i(t_r, t_s) \mid X_i = x, \Delta_i(t_r) = 1\}$ is the expected number of measurements from study time t_r to t_s for a subject i in group x , and $\nu_x(t_r) = \text{var}(K_i(t_r) \mid X_i = x, \Delta_i(t_r) = 1)$ be the variance of these measurement counts, $x = 0, 1$. Let $\mu_x = \mu(X_i = x)$ and let $V_x = \text{var}(Y_{ij} \mid X_i = x)$ be the true variance of responses for the subject in group x so that $V_x = \mu_x(1 - \mu_x)$, $x = 0, 1$. Additionally let $\pi(t_r)$ denote $P(\Delta_i(t_r) = 1)$, then further taking the expectation with respect to $K_i(t_r)$, X_i and $\Delta_i(t_r)$ we obtain

$$\mathcal{A}(t_r) = \pi(t_r) \begin{pmatrix} \kappa_1(t_r) \mu_1 P(X_i = 1) & \kappa_1(t_r) \mu_1 P(X_i = 1) \\ \kappa_1(t_r) \mu_1 P(X_i = 1) & \sum_{x=0}^1 \kappa_x(t_r) \mu_x P(X_i = x) \end{pmatrix}. \quad (\text{B.11})$$

Derivation of the $\mathcal{B}(t_r)$ Matrix

We have $\mathcal{B}(t_r) = \mathbb{E}\{\mathbf{U}_i(t_r; \boldsymbol{\beta}) \mathbf{U}_i'(t_r; \boldsymbol{\beta})\}$ which is

$$\mathbb{E} \left\{ \begin{array}{cc} \Delta_i(t_r) \sum_{j=1}^{K_i(t_r)} \sum_{k=1}^{K_i(t_r)} U_{ij1}(\beta) U_{ik1}(\beta) & \Delta_i(t_r) \sum_{j=1}^{K_i(t_r)} \sum_{k=1}^{K_i(t_r)} U_{ij1}(\beta) U_{ik2}(\beta) \\ \Delta_i(t_r) \sum_{j=1}^{K_i(t_r)} \sum_{k=1}^{K_i(t_r)} U_{ij1}(\beta) U_{ik2}(\beta) & \Delta_i(t_r) \sum_{j=1}^{K_i(t_r)} \sum_{k=1}^{K_i(t_r)} U_{ij2}(\beta) U_{ik2}(\beta) \end{array} \right\}. \quad (\text{B.12})$$

For the (1,1) entry of (B.12) we have

$$\begin{aligned} & \mathbb{E} \left\{ \Delta_i(t_r) \sum_{j=1}^{K_i(t_r)} \sum_{k=1}^{K_i(t_r)} U_{ij1}(\beta) U_{ik1}(\beta) \mid X_i, K_i(t_r), \Delta_i(t_r) \right\} \\ &= \Delta_i(t_r) \mathbb{E} \left\{ \sum_{j=1}^{K_i(t_r)} \sum_{k=1}^{K_i(t_r)} (Y_{ij} - \mu(X_i)) X_i (Y_{ik} - \mu(X_i)) \mid X_i, K_i(t_r), \Delta_i(t_r) \right\} \\ &= \Delta_i(t_r) \mathbb{E} \left\{ \sum_{j=1}^{K_i(t_r)} X_i U_{ij2}^2 + \sum_{\substack{j=1 \\ j \neq k}}^{K_i(t_r)} \sum_{k=1}^{K_i(t_r)} X_i (Y_{ij} - \mu(X_i)) (Y_{ik} - \mu(X_i)) \mid X_i, K_i(t_r), \Delta_i(t_r) \right\} \\ &= \Delta_i(t_r) \{K_i(t_r) \text{var}(Y_{ij} \mid X_i = 1) + K_i(t_r)(K_i(t_r) - 1) \text{cov}(Y_{ij}, Y_{ik} \mid X_i = 1)\} I(X_i = 1). \end{aligned}$$

Taking the expectation with respect to $K_i(t_r)$, X_i and $\Delta_i(t_r)$ gives

$$\pi(t_r) [\kappa_1(t_r) \text{var}(Y_{ij} \mid X_i = 1) + \{\nu_1(t_r) + \kappa_1^2(t_r) - \kappa_1(t_r)\} \text{cov}(Y_{ij}, Y_{ik} \mid X_i = 1)] P(X_i = 1).$$

For the (1,2) entry given X_i , $K_i(t_r)$ and $\Delta_i(t_r)$ we get

$$\Delta_i(t_r) \{K_i(t_r) \text{var}(Y_{ij} \mid X_i = 1) + K_i(t_r)(K_i(t_r) - 1) \text{cov}(Y_{ij}, Y_{ik} \mid X_i = 1)\} I(X_i = 1),$$

and overall we have

$$\pi(t_r) [\kappa_1(t_r) \text{var}(Y_{ij} \mid X_i = 1) + \{\nu_1(t_r) + \kappa_1^2(t_r) - \kappa_1(t_r)\} \text{cov}(Y_{ij}, Y_{ik} \mid X_i = 1)] P(X_i = 1).$$

Regarding the (2, 2) entry of (B.12) note that

$$\begin{aligned}
& \mathbb{E} \left\{ \Delta_i(t_r) \sum_{j=1}^{K_i(t_r)} \sum_{k=1}^{K_i(t_r)} U_{ij2}(\beta) U_{ik2}(\beta) \middle| X_i, K_i(t_r), \Delta_i(t_r) \right\} \\
&= \Delta_i(t_r) \mathbb{E} \left\{ \sum_{j=1}^{K_i(t_r)} U_{ij2}^2(\beta) + \sum_{j=1}^{K_i(t_r)} \sum_{\substack{k=1 \\ j \neq k}}^{K_i(t_r)} U_{ij2}(\beta) U_{ik2}(\beta) \middle| X_i, K_i(t_r), \Delta_i(t_r) \right\} \\
&= \Delta_i(t_r) \left\{ K_i(t_r) \text{var}(Y_{ij} \mid X_i) + \sum_{j=1}^{K_i(t_r)} \sum_{\substack{k=1 \\ j \neq k}}^{K_i(t_r)} \text{cov}(Y_{ij}, Y_{ik} \mid X_i) \right\} \\
&= \Delta_i(t_r) \{ K_i(t_r) \text{var}(Y_{ij} \mid X_i) + K_i(t_r) (K_i(t_r) - 1) \text{cov}(Y_{ij}, Y_{ik} \mid X_i) \}.
\end{aligned}$$

This has expectation

$$\begin{aligned}
& \sum_{x=0}^1 \left[\mathbb{E} \{ K_i(t_r) \mid X_i = x \} \text{var}(Y_{ij} \mid X_i = x) P(X_i = x) + \right. \\
& \quad \left. \mathbb{E} \{ K_i(t_r)(K_i(t_r) - 1) \mid X_i = x \} \text{cov}(Y_{ij}, Y_{ik} \mid X_i = x) P(X_i = x) \right] P(\Delta_i(t_r) = 1) \\
&= \pi(t_r) \sum_{x=0}^1 \left[\kappa_x(t_r) \text{var}(Y_{ij} \mid X_i = x) + \{ \nu_x(t_r) + \kappa_x^2(t_r) - \kappa_x(t_r) \} \text{cov}(Y_{ij}, Y_{ik} \mid X_i = x) \right] P(X_i = x).
\end{aligned}$$

If we let $\text{corr}(Y_{ij}, Y_{ik} \mid X_i = x) = \rho$ be an exchangeable correlation of responses within individuals we can express $\text{cov}(Y_{ij}, Y_{ik} \mid X_i = x)$ as $\text{cov}(Y_{ij}, Y_{ik} \mid X_i = x) = \rho \cdot \text{var}(Y_{ij} \mid X_i = x)$. Hence $\mathcal{B}(t_r)$ is

$$\pi(t_r) \begin{pmatrix} P(X_i = 1) \{ \kappa_1(t_r) V_1 + \delta_1(t_r) \rho V_1 \} & P(X_i = 1) \{ \kappa_1(t_r) V_1 + \delta_1(t_r) \rho V_1 \} \\ P(X_i = 1) \{ \kappa_1(t_r) V_1 + \delta_1(t_r) \rho V_1 \} & \sum_{x=0}^1 P(X_i = x) \{ \kappa_x(t_r) V_x + \delta_x(t_r) \rho V_x \} \end{pmatrix}, \quad (\text{B.13})$$

where $\delta_x(t_r) = \nu_x(t_r) + \kappa_x^2(t_r) - \kappa_x(t_r)$.

Derivation of the Covariance Matrix $\mathcal{B}(t_r, t_s)$

Recall that $\mathbf{U}_i(t_r; \boldsymbol{\beta})$ is given by

$$\begin{aligned}\mathbf{U}_i(t_r; \boldsymbol{\beta}) &= \Delta_i(t_r) \begin{pmatrix} \sum_{j=1}^{K_i(t_r)} X_i(Y_{ij} - \mu(X_i)) \\ \sum_{j=1}^{K_i(t_r)} (Y_{ij} - \mu(X_i)) \end{pmatrix} \\ &= \Delta_i(t_r) \sum_{j=1}^{K_i(t_r)} \begin{pmatrix} U_{ij1}(\boldsymbol{\beta}) \\ U_{ij2}(\boldsymbol{\beta}) \end{pmatrix}.\end{aligned}$$

Suppose $t_r < t_s$ so we have $K_i(t_r) \leq K_i(t_s)$ and

$$\begin{aligned}\mathcal{B}(t_r, t_s) &= \mathbb{E}\{\mathbf{U}_i(t_r; \boldsymbol{\beta})\mathbf{U}_i'(t_s; \boldsymbol{\beta})\} \\ &= \mathbb{E}\left\{ \begin{pmatrix} \Delta_i(t_r) \sum_{j=1}^{K_i(t_r)} \sum_{k=1}^{K_i(t_s)} U_{ij1}(\boldsymbol{\beta})U_{ik1}(\boldsymbol{\beta}) & \Delta_i(t_r) \sum_{j=1}^{K_i(t_r)} \sum_{k=1}^{K_i(t_s)} U_{ij1}(\boldsymbol{\beta})U_{ik2}(\boldsymbol{\beta}) \\ \Delta_i(t_r) \sum_{j=1}^{K_i(t_r)} \sum_{k=1}^{K_i(t_s)} U_{ij2}(\boldsymbol{\beta})U_{ik1}(\boldsymbol{\beta}) & \Delta_i(t_r) \sum_{j=1}^{K_i(t_r)} \sum_{k=1}^{K_i(t_s)} U_{ij2}(\boldsymbol{\beta})U_{ik2}(\boldsymbol{\beta}) \end{pmatrix} \right\}. \quad (\text{B.14})\end{aligned}$$

For the (1,1) entry of (B.14) we have

$$\begin{aligned}
& \mathbb{E} \left\{ \Delta_i(t_r) \sum_{j=1}^{K_i(t_r)} \sum_{k=1}^{K_i(t_s)} U_{ij1}(\beta) U_{ik1}(\beta) \mid X_i, K_i(t_r), K_i(t_s), \Delta_i(t_r) \right\} \\
&= \Delta_i(t_r) \mathbb{E} \left\{ \sum_{j=1}^{K_i(t_r)} \sum_{k=1}^{K_i(t_s)} (Y_{ij} - \mu(X_i)) X_i (Y_{ik} - \mu(X_i)) \mid X_i, K_i(t_r), K_i(t_s), \Delta_i(t_r) \right\} \\
&= \Delta_i(t_r) \mathbb{E} \left\{ \sum_{j=1}^{K_i(t_r)} X_i U_{ij2}^2 + \sum_{\substack{j=1 \\ j \neq k}}^{K_i(t_r)} \sum_{k=1}^{K_i(t_r)} X_i (Y_{ij} - \mu(X_i)) (Y_{ik} - \mu(X_i)) \right. \\
&\quad \left. + I(K_i(t_s) > K_i(t_r)) \sum_{j=1}^{K_i(t_r)} \sum_{k=K_i(t_r)+1}^{K_i(t_s)} X_i (Y_{ij} - \mu(X_i)) (Y_{ik} - \mu(X_i)) \mid X_i, K_i(t_r), K_i(t_s), \Delta_i(t_r) \right\} \\
&= \Delta_i(t_r) \{ K_i(t_r) \text{var}(Y_{ij} \mid X_i = 1) + K_i(t_r)(K_i(t_r) - 1) \text{cov}(Y_{ij}, Y_{ik} \mid X_i = 1) + \\
&\quad K_i(t_r)(K_i(t_s) - K_i(t_r)) \text{cov}(Y_{ij}, Y_{ik} \mid X_i = 1) \} I(X_i = 1) \\
&= \Delta_i(t_r) \{ K_i(t_r) \text{var}(Y_{ij} \mid X_i = 1) + K_i(t_r)(K_i(t_s) - 1) \text{cov}(Y_{ij}, Y_{ik} \mid X_i = 1) \} I(X_i = 1).
\end{aligned}$$

Taking the expectation with respect to $K_i(t_r)$, $K_i(t_s)$, X_i and $\Delta_i(t_r)$ gives

$$\pi(t_r) [\kappa_1(t_r) \text{var}(Y_{ij} \mid X_i = 1) + \{\nu_1(t_r) + \kappa_1^2(t_r) + \kappa_1(t_r)\kappa_1(t_r, t_s) - \kappa_1(t_r)\} \text{cov}(Y_{ij}, Y_{ik} \mid X_i = 1)] P(X_i = 1).$$

For the (1,2) entry of (B.14) given X_i , $K_i(t_r)$, $K_i(t_s)$ and $\Delta_i(t_r)$ we get

$$\Delta_i(t_r) \{ K_i(t_r) \text{var}(Y_{ij} \mid X_i = 1) + K_i(t_r)(K_i(t_s) - 1) \text{cov}(Y_{ij}, Y_{ik} \mid X_i = 1) \} I(X_i = 1),$$

and overall we have

$$\pi(t_r) [\kappa_1(t_r) \text{var}(Y_{ij} \mid X_i = 1) + \{\nu_1(t_r) + \kappa_1^2(t_r) + \kappa_1(t_r)\kappa_1(t_r, t_s) - \kappa_1(t_r)\} \text{cov}(Y_{ij}, Y_{ik} \mid X_i = 1)] P(X_i = 1).$$

Regarding the (2, 2) entry of (B.14) note that

$$\begin{aligned}
& \mathbb{E} \left\{ \Delta_i(t_r) \sum_{j=1}^{K_i(t_r)} \sum_{k=1}^{K_i(t_s)} U_{ij2}(\beta) U_{ik2}(\beta) \mid X_i, K_i(t_r), K_i(t_s), \Delta_i(t_r) \right\} \\
&= \Delta_i(t_r) \mathbb{E} \left\{ \sum_{j=1}^{K_i(t_r)} U_{ij2}^2(\beta) + \sum_{\substack{j=1 \\ j \neq k}}^{K_i(t_r)} \sum_{k=1}^{K_i(t_s)} U_{ij2}(\beta) U_{ik2}(\beta) \mid X_i, K_i(t_r), K_i(t_s), \Delta_i(t_r) \right\} \\
&= \Delta_i(t_r) \left\{ K_i(t_r) \text{var}(Y_{ij} \mid X_i) + \sum_{\substack{j=1 \\ j \neq k}}^{K_i(t_r)} \sum_{k=1}^{K_i(t_s)} \text{cov}(Y_{ij}, Y_{ik} \mid X_i) \right\} \\
&= \Delta_i(t_r) \left\{ K_i(t_r) \text{var}(Y_{ij} \mid X_i) + \sum_{\substack{j=1 \\ j \neq k}}^{K_i(t_r)} \sum_{k=1}^{K_i(t_r)} \text{cov}(Y_{ij}, Y_{ik} \mid X_i) + I(K_i(t_s) > K_i(t_r)) \sum_{j=1}^{K_i(t_r)} \sum_{k=K_i(t_r)+1}^{K_i(t_s)} \text{cov}(Y_{ij}, Y_{ik} \mid X_i) \right\} \\
&= \Delta_i(t_r) \left\{ K_i(t_r) \text{var}(Y_{ij} \mid X_i) + K_i(t_r)(K_i(t_r) - 1) \text{cov}(Y_{ij}, Y_{ik} \mid X_i) + K_i(t_r)(K_i(t_s) - K_i(t_r)) \text{cov}(Y_{ij}, Y_{ik} \mid X_i) \right\} \\
&= \Delta_i(t_r) \{ K_i(t_r) \text{var}(Y_{ij} \mid X_i) + K_i(t_r) (K_i(t_s) - 1) \text{cov}(Y_{ij}, Y_{ik} \mid X_i) \}.
\end{aligned}$$

This has expectation

$$\begin{aligned}
& \sum_{x=0}^1 \left[\mathbb{E}\{K_i(t_r) \mid X_i = x\} \text{var}(Y_{ij} \mid X_i = x) P(X_i = x) + \right. \\
& \quad \left. \mathbb{E}\{K_i(t_r)(K_i(t_s) - 1) \mid X_i = x\} \text{cov}(Y_{ij}, Y_{ik} \mid X_i = x) P(X_i = x) \right] P(\Delta_i(t_r) = 1) \\
&= \pi(t_r) \sum_{x=0}^1 \left[\kappa_x(t_r) \text{var}(Y_{ij} \mid X_i = x) + \{\nu_x(t_r) + \kappa_x^2(t_r) + \kappa_x(t_r)\kappa_x(t_r, t_s) - \kappa_x(t_r)\} \text{cov}(Y_{ij}, Y_{ik} \mid X_i = x) \right] P(X_i = x).
\end{aligned}$$

Hence $\mathcal{B}(t_r, t_s)$ is

$$\pi(t_r) \begin{pmatrix} P(X_i = 1) \{ \kappa_1(t_r) V_1 + \delta_1(t_r, t_s) \rho V_1 \} & P(X_i = 1) \{ \kappa_1(t_r) V_1 + \delta_1(t_r, t_s) \rho V_1 \} \\ P(X_i = 1) \{ \kappa_1(t_r) V_1 + \delta_1(t_r, t_s) \rho V_1 \} & \sum_{x=0}^1 P(X_i = x) \{ \kappa_x(t_r) V_x + \delta_x(t_r, t_s) \rho V_x \} \end{pmatrix}, \quad (\text{B.15})$$

where $\delta_x(t_r, t_s) = \nu_x(t_r) + \kappa_x^2(t_r) + \kappa_x(t_r)\kappa_x(t_r, t_s) - \kappa_x(t_r)$.

B.5 Derivations for Design of a Two-Arm Clinical Trial

Simplification of the Asymptotic Variance of $\hat{\beta}_1(t_r)$

With 50:50 randomization we can write (B.11) as

$$\mathcal{A}(t_r) = \pi(t_r) \begin{pmatrix} \frac{\kappa_1(t_r)\mu_1}{2} & \frac{\kappa_1(t_r)\mu_1}{2} \\ \frac{\kappa_1(t_r)\mu_1}{2} & \frac{\kappa_0(t_r)\mu_0}{2} + \frac{\kappa_1(t_r)\mu_1}{2} \end{pmatrix},$$

and (B.13) as

$$\mathcal{B}(t_r) = \pi(t_r) \begin{pmatrix} \frac{1}{2} \{\kappa_1(t_r)V_1 + \delta_1(t_r)\rho V_1\} & \frac{1}{2} \{\kappa_1(t_r)V_1 + \delta_1(t_r)\rho V_1\} \\ \frac{1}{2} \{\kappa_1(t_r)V_1 + \delta_1(t_r)\rho V_1\} & \frac{1}{2} \{\kappa_0(t_r)V_0 + \delta_0(t_r)\rho V_0 + \kappa_1(t_r)V_1 + \delta_1(t_r)\rho V_1\} \end{pmatrix},$$

and (B.15) as

$$\mathcal{B}(t_r, t_s) = \pi(t_r) \begin{pmatrix} \frac{1}{2} \{\kappa_1(t_r)V_1 + \delta_1(t_r, t_s)\rho V_1\} & \frac{1}{2} \{\kappa_1(t_r)V_1 + \delta_1(t_r, t_s)\rho V_1\} \\ \frac{1}{2} \{\kappa_1(t_r)V_1 + \delta_1(t_r, t_s)\rho V_1\} & \frac{1}{2} \{\kappa_0(t_r)V_0 + \delta_0(t_r, t_s)\rho V_0 + \kappa_1(t_r)V_1 + \delta_1(t_r, t_s)\rho V_1\} \end{pmatrix}.$$

Recall that

$$\begin{aligned} M_2(t_r, t_s) &= \text{cov}(\bar{U}_i(t_r; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha}), \bar{U}_i(t_s; \beta_1, \boldsymbol{\theta}, \boldsymbol{\alpha})) \\ &= \mathcal{B}_{\beta_1\beta_1}(t_r, t_s) + \mathcal{A}_{\beta_1\boldsymbol{\theta}}(t_r)\mathcal{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(t_r)\mathcal{B}_{\boldsymbol{\theta}\boldsymbol{\theta}}(t_r, t_s)[\mathcal{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(t_s)]'[\mathcal{A}_{\beta_1\boldsymbol{\theta}}(t_s)]' \\ &\quad - \mathcal{B}_{\beta_1\boldsymbol{\theta}}(t_r, t_s)[\mathcal{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(t_s)]'[\mathcal{A}_{\beta_1\boldsymbol{\theta}}(t_s)]' - \mathcal{A}_{\beta_1\boldsymbol{\theta}}(t_r)\mathcal{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(t_r)\mathcal{B}_{\boldsymbol{\theta}\beta_1}(t_r, t_s). \end{aligned}$$

Then letting $a_x(t_r, t_s) = \frac{1}{2} \{\kappa_x(t_r)V_x + \delta_x(t_r, t_s)\rho V_x\}$, $b_x(t_r) = \kappa_x(t_r)\mu_x/2$, $x = 0, 1$,

$$\begin{aligned} M_2(t_r, t_s) &= \pi(t_r) \left\{ a_1(t_r, t_s) + \frac{b_1(t_r)(a_0(t_r, t_s) + a_1(t_r, t_s))b_1(t_s)}{\{b_0(t_r) + b_1(t_r)\}\{b_0(t_s) + b_1(t_s)\}} - \frac{a_1(t_r, t_s)b_1(t_s)}{b_0(t_s) + b_1(t_s)} - \frac{a_1(t_r, t_s)b_1(t_r)}{b_0(t_r) + b_1(t_r)} \right\} \\ &= \pi(t_r) \frac{b_0(t_r)b_0(t_s)a_1(t_r, t_s) + b_1(t_r)b_1(t_s)a_0(t_r, t_s)}{\{b_0(t_r) + b_1(t_r)\}\{b_0(t_s) + b_1(t_s)\}}, \end{aligned}$$

and substituting the parameters for the constants further gives $M_2(t_r, t_s)$ as

$$\pi(t_r) \frac{\kappa_0(t_r)\kappa_0(t_s)\mu_0^2 V_1\{\kappa_1(t_r) + \delta_1(t_r, t_s)\rho\} + \kappa_1(t_r)\kappa_1(t_s)\mu_1^2 V_0\{\kappa_0(t_r) + \delta_0(t_r, t_s)\rho\}}{2\{\kappa_0(t_r)\mu_0 + \kappa_1(t_r)\mu_1\}\{\kappa_0(t_s)\mu_0 + \kappa_1(t_s)\mu_1\}}.$$

Similarly we have

$$M_1(t_r) = \pi(t_r) \frac{\kappa_0^2(t_r)\mu_0^2 V_1\{\kappa_1(t_r) + \delta_1(t_r)\rho\} + \kappa_1^2(t_r)\mu_1^2 V_0\{\kappa_0(t_r) + \delta_0(t_r)\rho\}}{2\{\kappa_0(t_r)\mu_0 + \kappa_1(t_r)\mu_1\}^2}.$$

Simplification of the Asymptotic Variance of $\hat{\beta}_1(t_r)$

Note that

$$\begin{aligned} \mathcal{A}^{-1}(t_r) &= \pi(t_r)^{-1} \begin{pmatrix} b_1(t_r) & b_1(t_r) \\ b_1(t_r) & b_0(t_r) + b_1(t_r) \end{pmatrix}^{-1} \\ &= \frac{\pi(t_r)^{-1}}{b_0(t_r)b_1(t_r) + b_1(t_r)^2 - b_1(t_r)^2} \begin{pmatrix} b_0(t_r) + b_1(t_r) & -b_1(t_r) \\ -b_1(t_r) & b_1(t_r) \end{pmatrix} \\ &= \pi(t_r)^{-1} \begin{pmatrix} \frac{1}{b_0(t_r)} + \frac{1}{b_1(t_r)} & -\frac{1}{b_0(t_r)} \\ -\frac{1}{b_0(t_r)} & \frac{1}{b_0(t_r)} \end{pmatrix}, \end{aligned}$$

which gives

$$\mathcal{A}^{-1}(t_r) = \pi(t_r)^{-1} \begin{pmatrix} \frac{2}{\kappa_0(t_r)\mu_0} + \frac{2}{\kappa_1(t_r)\mu_1} & -\frac{2}{\kappa_0(t_r)\mu_0} \\ -\frac{2}{\kappa_0(t_r)\mu_0} & \frac{2}{\kappa_0(t_r)\mu_0} \end{pmatrix}.$$

Then

$$\begin{aligned} \Omega(t_r) &= [\mathcal{A}^{-1}(t_r)] \mathcal{B}(t_r) [\mathcal{A}^{-1}(t_r)]' \\ &= \frac{1}{\pi(t_r)} \begin{pmatrix} \frac{2V_0}{\kappa_0^2(t_r)\mu_0^2} [\kappa_0(t_r) + \delta_0(t_r)\rho] + \frac{2V_1}{\kappa_1^2(t_r)\mu_1^2} [\kappa_1(t_r) + \delta_1(t_r)\rho] & -\frac{2V_0}{\kappa_0^2(t_r)\mu_0^2} [\kappa_0(t_r) + \delta_0(t_r)\rho] \\ -\frac{2V_0}{\kappa_0^2(t_r)\mu_0^2} [\kappa_0(t_r) + \delta_0(t_r)\rho] & \frac{2V_0}{\kappa_0^2(t_r)\mu_0^2} [\kappa_0(t_r) + \delta_0(t_r)\rho] \end{pmatrix}. \end{aligned}$$

Thus the asymptotic variance of $\widehat{\beta}_1(t_r)$ is

$$\frac{1}{n_r} \pi(t_r)^{-1} \left\{ \frac{2V_0}{\kappa_0(t_r)\mu_0^2} \left(1 + \frac{\delta_0(t_r)\rho}{\kappa_0(t_r)} \right) + \frac{2V_1}{\kappa_1(t_r)\mu_1^2} \left(1 + \frac{\delta_1(t_r)\rho}{\kappa_1(t_r)} \right) \right\}. \quad (\text{B.16})$$

B.6 Derivations of $\kappa_x(t_r)$, $\nu_x(t_r)$, and $\kappa_x(t_r, t_s)$

Suppose K is the number of transfusions per individual over a period of time of duration τ , $K \mid u \sim \text{Poisson}(u\lambda\tau)$, and $L \mid U = u \sim \text{Exp}(u\lambda)$ and the random effect U follows the Gamma distribution with $\mathbb{E}(U) = 1$ and $\text{var}(U) = \phi$. We have

$$f(l; \lambda, \phi) = \int_0^\infty P(l \mid u; \lambda) dG(u; \phi) = \int_0^\infty u \lambda \exp(-u \lambda l) \frac{u^{\phi^{-1}-1} e^{-u/\phi}}{\Gamma(\phi^{-1}) \phi^{\phi^{-1}}} du = \lambda \left(\frac{1}{\lambda l \phi + 1} \right)^{\phi^{-1}+1}, \quad (\text{B.17})$$

and

$$\begin{aligned} dG(u \mid l; \lambda, \phi) &= \frac{P(l \mid u; \lambda) dG(u; \phi)}{\int_0^\infty P(l \mid u; \lambda) dG(u; \phi)} = \frac{u \lambda \exp(-u \lambda l) \frac{u^{\phi^{-1}-1} e^{-u/\phi}}{\Gamma(\phi^{-1}) \phi^{\phi^{-1}}}}{\int_0^\infty u \lambda \exp(-u \lambda l) \frac{u^{\phi^{-1}-1} e^{-u/\phi}}{\Gamma(\phi^{-1}) \phi^{\phi^{-1}}} du} \\ &= \frac{u^{\phi^{-1}} \exp\left(-u / \frac{\phi}{\lambda l + 1}\right)}{\Gamma(\phi^{-1} + 1) \left(\frac{\phi}{\lambda l + 1}\right)^{\phi^{-1}+1}}, \end{aligned}$$

which gives

$$U \mid L = l \sim \text{Gamma} \left(\text{shape} = \phi^{-1} + 1, \text{scale} = \frac{\phi}{\lambda l + 1} \right).$$

To compute the $\kappa_x(t_r)$ we average over the random effect U and the recruitment time L as

$$\begin{aligned} \kappa_x(t_r) &= \mathbb{E}\{K_i(t_r) \mid X_i = x, \Delta_i(t_r) = 1\} \\ &= \mathbb{E}_L \left\{ \mathbb{E}_U \left\{ \mathbb{E}\{K_i(t_r) \mid X_i = x, \Delta_i(t_r) = 1, \right. \right. \right. \\ &\quad \left. \left. \left. U_i = u, L_i = l \right\} \mid \Delta_i(t_r) = 1, L_i < \min(t_r, \tau_A) \right\} \mid \Delta_i(t_r) = 1, L_i < \min(t_r, \tau_A) \right\}. \end{aligned}$$

Note that

$$g_0(t_1, t_2) = \int_{t_1}^{t_2} f(l) dl = \int_{t_1}^{t_2} \frac{\lambda}{(\lambda\phi l + 1)^{\phi^{-1}+1}} dl = (\lambda\phi t_1 + 1)^{-\phi^{-1}} - (\lambda\phi t_2 + 1)^{-\phi^{-1}}, \quad (\text{B.18})$$

$$\begin{aligned} g_1(t_1, t_2) &= \int_{t_1}^{t_2} l f(l) dl = \int_{t_1}^{t_2} l \frac{\lambda}{(\lambda\phi l + 1)^{\phi^{-1}+1}} dl = \lambda \int_{t_1}^{t_2} l d \left\{ \frac{1}{-\lambda} (\lambda\phi l + 1)^{-\phi^{-1}} \right\} \\ &= \left\{ -l(\lambda\phi l + 1)^{-\phi^{-1}} + \frac{1}{\lambda\phi - \lambda} (\lambda\phi l + 1)^{-\phi^{-1}+1} \right\} \Big|_{t_1}^{t_2} \\ &= \frac{1}{\lambda\phi - \lambda} \left\{ (\lambda\phi t_2 + 1)^{-\phi^{-1}+1} - (\lambda\phi t_1 + 1)^{-\phi^{-1}+1} \right\} - t_2(\lambda\phi t_2 + 1)^{-\phi^{-1}} + t_1(\lambda\phi t_1 + 1)^{-\phi^{-1}}, \end{aligned} \quad (\text{B.19})$$

$$\begin{aligned} g_2(t_1, t_2) &= \int_{t_1}^{t_2} l^2 f(l) dl = \int_{t_1}^{t_2} l^2 \frac{\lambda}{(\lambda\phi l + 1)^{\phi^{-1}+1}} dl = \lambda \int_{t_1}^{t_2} l^2 d \left\{ \frac{1}{-\lambda} (\lambda\phi l + 1)^{-\phi^{-1}} \right\} \\ &= -l^2(\lambda\phi l + 1)^{-\phi^{-1}} \Big|_{t_1}^{t_2} + 2 \int_{t_1}^{t_2} l(\lambda\phi l + 1)^{-\phi^{-1}} dl \\ &= -l^2(\lambda\phi l + 1)^{-\phi^{-1}} \Big|_{t_1}^{t_2} + 2 \int_{t_1}^{t_2} l d \left\{ \frac{1}{-\lambda(1-\phi)} (\lambda\phi l + 1)^{-\phi^{-1}+1} \right\} \\ &= \left\{ -l^2(\lambda\phi l + 1)^{-\phi^{-1}} - \frac{2l}{\lambda(1-\phi)} (\lambda\phi l + 1)^{-\phi^{-1}+1} - \frac{2}{\lambda^2(1-2\phi)(1-\phi)} (\lambda\phi l + 1)^{-\phi^{-1}+2} \right\} \Big|_{t_1}^{t_2} \\ &= -t_2^2(\lambda\phi t_2 + 1)^{-\phi^{-1}} + \frac{2t_2}{\lambda(\phi-1)} (\lambda\phi t_2 + 1)^{-\phi^{-1}+1} - \frac{2}{\lambda^2(1-2\phi)(1-\phi)} (\lambda\phi t_2 + 1)^{-\phi^{-1}+2} \\ &\quad + t_1^2(\lambda\phi t_1 + 1)^{-\phi^{-1}} - \frac{2t_1}{\lambda(\phi-1)} (\lambda\phi t_1 + 1)^{-\phi^{-1}+1} + \frac{2}{\lambda^2(1-2\phi)(1-\phi)} (\lambda\phi t_1 + 1)^{-\phi^{-1}+2}. \end{aligned} \quad (\text{B.20})$$

Hence

$$\mathbb{E}(l \mid t_1 < l < t_2) = \frac{\int_{t_1}^{t_2} l f(l) dl}{\int_{t_1}^{t_2} f(l) dl} = \frac{g_1(t_1, t_2)}{g_0(t_1, t_2)}, \quad (\text{B.21})$$

$$\mathbb{E}(l^2 \mid t_1 < l < t_2) = \frac{\int_{t_1}^{t_2} l^2 f(l) dl}{\int_{t_1}^{t_2} f(l) dl} = \frac{g_2(t_1, t_2)}{g_0(t_1, t_2)}. \quad (\text{B.22})$$

Additionally,

$$\begin{aligned}
\mathbb{E}\{u \mid l < t\} &= \frac{\int_0^t \int_0^\infty u f(u, l) du dl}{\int_0^t \int_0^\infty f(u, l) du dl} = \frac{\int_0^t \int_0^\infty u f(u \mid l) f(l) du dl}{\int_0^t \int_0^\infty f(u \mid l) f(l) du dl} \\
&= \frac{\int_0^t f(l) \left\{ \int_0^\infty u f(u \mid l) du \right\} dl}{\int_0^t f(l) \left\{ \int_0^\infty f(u \mid l) du \right\} dl} = \frac{\int_0^t \frac{1+\phi}{\lambda\phi l+1} f(l) dl}{\int_0^t f(l) dl} \\
&= \frac{1 - (\lambda\phi t + 1)^{-\phi^{-1}-1}}{1 - (\lambda\phi t + 1)^{-\phi^{-1}}}, \tag{B.23}
\end{aligned}$$

and

$$\mathbb{E}\{u^2 \mid l < t\} = \frac{\int_0^t f(l) \left\{ \int_0^\infty u^2 f(u \mid l) du \right\} dl}{\int_0^t f(l) dl} \tag{B.24}$$

$$\begin{aligned}
&= \frac{\int_0^t f(l) \left\{ (\phi^{-1} + 1) \left(\frac{\phi}{\lambda\phi l+1} \right)^2 + \left(\frac{1+\phi}{\lambda\phi l+1} \right)^2 \right\} dl}{\int_0^t f(l) dl} \\
&= \frac{(\phi + 1)(2\phi + 1) \int_0^t f(l) (\lambda\phi l + 1)^{-2} dl}{1 - (\lambda\phi t + 1)^{-\phi^{-1}}} \\
&= (1 + \phi) \frac{1 - (\lambda\phi t + 1)^{-\phi^{-1}-2}}{1 - (\lambda\phi t + 1)^{-\phi^{-1}}}. \tag{B.25}
\end{aligned}$$

With (B.18)-(B.24) we consider two cases for the study time t_r where 1) $0 < t_r \leq \tau_F$, 2) $\tau_F < t_r \leq t_R$.

For 1) $0 < t_r \leq \tau_F$,

$$\begin{aligned}
\kappa_x(t_r) &= \mathbb{E}\{K_i(t_r) \mid X_i = x, \Delta_i(t_r) = 1\} \\
&= \mathbb{E}_L \left\{ \mathbb{E}_U \left\{ \mathbb{E}\{K_i(t_r) \mid X_i = x, \Delta_i(t_r) = 1, \right. \right. \\
&\quad \left. \left. U_i = u, L_i = l\} \mid \Delta_i(t_r) = 1, L_i < \min(t_r, \tau_A) \right\} \mid \Delta_i(t_r) = 1, L_i < \min(t_r, \tau_A) \right\}.
\end{aligned}$$

We omit the conditions $X_i = x$ and $\Delta_i(t_r) = 1$ in the expectations for the ease of the

notation thus

$$\begin{aligned}
\kappa_x(t_r) &= \mathbb{E}_L \{ \mathbb{E}_U \{ \mathbb{E} \{ K_i(t_r) \mid U_i = u, L_i = l \} \mid l < \min(t_r, \tau_A) \} \mid l < \min(t_r, \tau_A) \} \\
&= \mathbb{E}_L \{ \mathbb{E}_U \{ u \lambda(t_r - l) \mid l < \min(t_r, \tau_A) \} \mid l < \min(t_r, \tau_A) \} \\
&= \lambda \{ t_r - \mathbb{E} \{ l \mid l < \min(t_r, \tau_A) \} \} \mathbb{E} \{ u \mid l < \min(t_r, \tau_A) \} \\
&= \lambda \left\{ t_r - \frac{g_1(0, \min(t_r, \tau_A))}{g_0(0, \min(t_r, \tau_A))} \right\} \mathbb{E} \{ u \mid l < \min(t_r, \tau_A) \} ,
\end{aligned}$$

and

$$\begin{aligned}
\nu_x(t_r) &= \text{var} \{ K_i(t_r) \mid X_i = x, \Delta_i(t_r) = 1 \} \\
&= \mathbb{E} \{ K_i^2(t_r) \mid X_i = x, \Delta_i(t_r) = 1 \} - \kappa_x^2(t_r) \\
&= \mathbb{E}_L \{ \mathbb{E}_U \{ \mathbb{E} \{ K_i^2(t_r) \mid U_i = u, L_i = l \} \mid l < \min(t_r, \tau_A) \} \mid l < \min(t_r, \tau_A) \} - \kappa_x^2(t_r) \\
&= \mathbb{E}_L \{ \mathbb{E}_U \{ \text{var} \{ K_i(t_r) \mid U_i = u, L_i = l \} \\
&\quad + \mathbb{E}^2 \{ K_i(t_r) \mid U_i = u, L_i = l \} \mid l < \min(t_r, \tau_A) \} \mid l < \min(t_r, \tau_A) \} - \kappa_x^2(t_r) \\
&= \mathbb{E}_L \{ \mathbb{E}_U \{ u \lambda(t_r - l) + u^2 \lambda^2(t_r - l)^2 \mid l < \min(t_r, \tau_A) \} \mid l < \min(t_r, \tau_A) \} - \kappa_x^2(t_r) \\
&= \kappa_x(t_r) - \kappa_x^2(t_r) + \lambda^2 \{ t_r^2 + \mathbb{E} \{ l^2 \mid l < \min(t_r, \tau_A) \} - 2t_r \mathbb{E} \{ l \mid l < \min(t_r, \tau_A) \} \} \mathbb{E} \{ u^2 \mid l < \min(t_r, \tau_A) \} \\
&= \kappa_x(t_r) - \kappa_x^2(t_r) + \lambda^2 \left\{ t_r^2 + \frac{g_2(0, \min(t_r, \tau_A))}{g_0(0, \min(t_r, \tau_A))} - 2t_r \frac{g_1(0, \min(t_r, \tau_A))}{g_0(0, \min(t_r, \tau_A))} \right\} \mathbb{E} \{ u^2 \mid l < \min(t_r, \tau_A) \} .
\end{aligned}$$

For 2) $\tau_F < t_r \leq t_R$ we have

$$\begin{aligned}
\kappa_x(t_r) &= \mathbb{E} \{ K_i(t_r) \mid X_i = x, \Delta_i(t_r) = 1 \} \\
&= \mathbb{E}_L \{ \mathbb{E}_U \{ \mathbb{E} \{ K_i(t_r) \mid U_i = u, L_i = l \} \mid l < \min(t_r, \tau_A) \} \mid l < \min(t_r, \tau_A) \} \\
&= \mathbb{E}_L \{ \mathbb{E}_U \{ u \lambda \min(t_r - l, \tau_F) \mid l < \min(t_r, \tau_A) \} \mid l < \min(t_r, \tau_A) \} \\
&= \lambda \mathbb{E} \{ u \mid l < \min(t_r, \tau_A) \} \mathbb{E}_L \{ \min(t_r - l, \tau_F) \mid l < \min(t_r, \tau_A) \} \\
&= \lambda \mathbb{E} \{ u \mid l < \min(t_r, \tau_A) \} \left\{ \frac{\int_{t_r - \tau_F}^{\min(t_r, \tau_A)} (t_r - l) f(l) dl + \int_0^{t_r - \tau_F} \tau_F f(l) dl}{g_0(0, \min(t_r, \tau_A))} \right\} \\
&= \lambda \mathbb{E} \{ u \mid l < \min(t_r, \tau_A) \} \left\{ t_r \frac{g_0(t_r - \tau_F, \min(t_r, \tau_A))}{g_0(0, \min(t_r, \tau_A))} - \frac{g_1(t_r - \tau_F, \min(t_r, \tau_A))}{g_0(0, \min(t_r, \tau_A))} + \tau_F \frac{g_0(0, t_r - \tau_F)}{g_0(0, \min(t_r, \tau_A))} \right\} ,
\end{aligned}$$

and

$$\begin{aligned}
\nu_x(t_r) &= \text{var}\{K_i(t_r) \mid X_i = x, \Delta_i(t_r) = 1\} \\
&= \mathbb{E}\{K_i^2(t_r) \mid X_i = x, \Delta_i(t_r) = 1\} - \kappa_x^2(t_r) \\
&= \mathbb{E}_L \left\{ \mathbb{E}_U \left\{ \mathbb{E} \{K_i^2(t_r) \mid U_i = u, L_i = l\} \mid l < \min(t_r, \tau_A) \right\} \mid l < \min(t_r, \tau_A) \right\} - \kappa_x^2(t_r) \\
&= \mathbb{E}_L \left\{ \mathbb{E}_U \left\{ \text{var}\{K_i(t_r) \mid U_i = u, L_i = l\} \right. \right. \\
&\quad \left. \left. + \mathbb{E}^2\{K_i(t_r) \mid U_i = u, L_i = l\} \mid l < \min(t_r, \tau_A) \right\} \mid l < \min(t_r, \tau_A) \right\} - \kappa_x^2(t_r) \\
&= \mathbb{E}_L \left\{ \mathbb{E}_U \left\{ u\lambda \min(t_r - l, \tau_F) + u^2 \lambda^2 (\min(t_r - l, \tau_F))^2 \mid l < \min(t_r, \tau_A) \right\} \mid l < \min(t_r, \tau_A) \right\} \\
&\quad - \kappa_x^2(t_r) \\
&= \kappa_x(t_r) + \lambda^2 \mathbb{E} \{u^2 \mid l < \min(t_r, \tau_A)\} \mathbb{E}_L \left\{ (\min(t_r - l, \tau_F))^2 \mid l < \min(t_r, \tau_A) \right\} - \kappa_x^2(t_r) \\
&= \kappa_x(t_r) + \lambda^2 \mathbb{E} \{u^2 \mid l < \min(t_r, \tau_A)\} \left\{ \frac{\int_{t_r - \tau_F}^{\min(t_r, \tau_A)} (t_r - l)^2 f(l) \, dl + \int_0^{t_r - \tau_F} \tau_F^2 f(l) \, dl}{g_0(0, \min(t_r, \tau_A))} \right\} - \kappa_x^2(t_r) \\
&= \kappa_x(t_r) + \lambda^2 \mathbb{E} \{u^2 \mid l < \min(t_r, \tau_A)\} \left\{ \frac{1}{g_0(0, \min(t_r, \tau_A))} \left\{ \int_{t_r - \tau_F}^{\min(t_r, \tau_A)} t_r^2 f(l) \, dl - \right. \right. \\
&\quad \left. \left. \int_{t_r - \tau_F}^{\min(t_r, \tau_A)} 2t_r l f(l) \, dl + \int_{t_r - \tau_F}^{\min(t_r, \tau_A)} l^2 f(l) \, dl + \int_0^{t_r - \tau_F} \tau_F^2 f(l) \, dl \right\} \right\} - \kappa_x^2(t_r) \\
&= \kappa_x(t_r) + \lambda^2 \mathbb{E} \{u^2 \mid l < \min(t_r, \tau_A)\} \left\{ \frac{1}{g_0(0, \min(t_r, \tau_A))} \left\{ t_r^2 g_0(t_r - \tau_F, \min(t_r, \tau_A)) \right. \right. \\
&\quad \left. \left. - 2t_r g_1(t_r - \tau_F, \min(t_r, \tau_A)) + g_2(t_r - \tau_F, \min(t_r, \tau_A)) + \tau_F^2 g_0(0, t_r - \tau_F) \right\} \right\} - \kappa_x^2(t_r) \\
&= \kappa_x(t_r) + \lambda^2 \mathbb{E} \{u^2 \mid l < \min(t_r, \tau_A)\} \left\{ t_r^2 \frac{g_0(t_r - \tau_F, \min(t_r, \tau_A))}{g_0(0, \min(t_r, \tau_A))} - 2t_r \frac{g_1(t_r - \tau_F, \min(t_r, \tau_A))}{g_0(0, \min(t_r, \tau_A))} \right. \\
&\quad \left. + \frac{g_2(t_r - \tau_F, \min(t_r, \tau_A))}{g_0(0, \min(t_r, \tau_A))} + \tau_F^2 \frac{g_0(0, t_r - \tau_F)}{g_0(0, \min(t_r, \tau_A))} \right\} - \kappa_x^2(t_r).
\end{aligned}$$

To compute $\kappa_x(t_r, t_s)$ we consider three cases 1) $0 < t_r < t_s \leq \tau_F$, 2) $0 < t_r < \tau_F < t_s \leq t_R$, 3) $\tau_F \leq t_r < t_s \leq t_R$.

For 1) $0 < t_r < t_s \leq \tau_F$,

$$\begin{aligned}
\kappa_x(t_r, t_s) &= \mathbb{E}\{K_i(t_r, t_s) \mid X_i = x, \Delta_i(t_r) = 1\} \\
&= \mathbb{E}_L \{ \mathbb{E}_U \{ \mathbb{E} \{ K_i(t_r, t_s) \mid U_i = u, L_i = l \} \mid l < \min(t_r, \tau_A) \} \mid l < \min(t_r, \tau_A) \} \\
&= \mathbb{E}_L \{ \mathbb{E}_U \{ u \lambda (t_s - t_r) \mid l < \min(t_r, \tau_A) \} \mid l < \min(t_r, \tau_A) \} \\
&= \lambda (t_s - t_r) \mathbb{E} \{ u \mid l < \min(t_r, \tau_A) \} .
\end{aligned}$$

For 2) $0 < t_r < \tau_F < t_s \leq t_R$,

$$\begin{aligned}
\kappa_x(t_r, t_s) &= \mathbb{E}\{K_i(t_r, t_s) \mid X_i = x, \Delta_i(t_r) = 1\} \\
&= \mathbb{E}_L \{ \mathbb{E}_U \{ \mathbb{E} \{ K_i(t_r, t_s) \mid U_i = u, L_i = l \} \mid l < \min(t_r, \tau_A) \} \mid l < \min(t_r, \tau_A) \} \\
&= \mathbb{E}_L \{ \mathbb{E}_U \{ u \lambda \min(t_s - t_r, l + \tau_F - t_r) \mid l < \min(t_r, \tau_A) \} \mid l < \min(t_r, \tau_A) \} \\
&= \lambda \mathbb{E} \{ u \mid l < \min(t_r, \tau_A) \} \mathbb{E}_L \{ \min(t_s - t_r, l + \tau_F - t_r) \mid l < \min(t_r, \tau_A) \} \\
&= \lambda \mathbb{E} \{ u \mid l < \min(t_r, \tau_A) \} \frac{\int_{t_s - \tau_F}^{\min(t_r, \tau_A)} (t_s - t_r) f(l) \, dl + \int_0^{t_s - \tau_F} (l + \tau_F - t_r) f(l) \, dl}{g_0(0, \min(t_r, \tau_A))} \\
&= \lambda \mathbb{E} \{ u \mid l < \min(t_r, \tau_A) \} \times \\
&\quad \frac{(t_s - t_r) g_0(t_s - \tau_F, \min(t_r, \tau_A)) + g_1(0, t_s - \tau_F) + (\tau_F - t_r) g_0(0, t_s - \tau_F)}{g_0(0, \min(t_r, \tau_A))} \\
&= \lambda \mathbb{E} \{ u \mid l < \min(t_r, \tau_A) \} \left\{ (t_s - t_r) \frac{g_0(t_s - \tau_F, \min(t_r, \tau_A))}{g_0(0, \min(t_r, \tau_A))} + \frac{g_1(0, t_s - \tau_F)}{g_0(0, \min(t_r, \tau_A))} + \right. \\
&\quad \left. (\tau_F - t_r) \frac{g_0(0, t_s - \tau_F)}{g_0(0, \min(t_r, \tau_A))} \right\} .
\end{aligned}$$

For 3) $\tau_F \leq t_r < t_s \leq t_R$,

$$\begin{aligned}
\kappa_x(t_r, t_s) &= \mathbb{E}\{K_i(t_r, t_s) \mid X_i = x, \Delta_i(t_r) = 1\} \\
&= \mathbb{E}_L \{ \mathbb{E}_U \{ \mathbb{E} \{ K_i(t_r, t_s) \mid U_i = u, L_i = l \} \mid l < \min(t_r, \tau_A) \} \mid l < \min(t_r, \tau_A) \} \\
&= \mathbb{E}_L \{ \mathbb{E}_U \{ u \lambda \max(\min(t_s - t_r, l + \tau_F - t_r), 0) \mid l < \min(t_r, \tau_A) \} \mid l < \min(t_r, \tau_A) \} \\
&= \lambda \mathbb{E} \{ u \mid l < \min(t_r, \tau_A) \} \mathbb{E}_L \{ \max(\min(t_s - t_r, l + \tau_F - t_r), 0) \mid l < \min(t_r, \tau_A) \} \\
&= \lambda \mathbb{E} \{ u \mid l < \min(t_r, \tau_A) \} \left\{ \frac{\int_{t_s - \tau_F}^{\min(t_r, \tau_A)} (t_s - t_r) f(l) dl + \int_0^{t_s - \tau_F} (l + \tau_F - t_r) f(l) dl}{g_0(0, \min(t_r, \tau_A))} \right\} \\
&= \lambda \mathbb{E} \{ u \mid l < \min(t_r, \tau_A) \} \times \\
&\quad \frac{(t_s - t_r)g_0(t_s - \tau_F, \min(t_r, \tau_A)) + g_1(0, t_s - \tau_F) + (\tau_F - t_r)g_0(0, t_s - \tau_F)}{g_0(0, \min(t_r, \tau_A))} \\
&= \lambda \mathbb{E} \{ u \mid l < \min(t_r, \tau_A) \} \times \\
&\quad \left\{ (t_s - t_r) \frac{g_0(t_s - \tau_F, \min(t_r, \tau_A))}{g_0(0, \min(t_r, \tau_A))} + \frac{g_1(0, t_s - \tau_F)}{g_0(0, \min(t_r, \tau_A))} + (\tau_F - t_r) \frac{g_0(0, t_s - \tau_F)}{g_0(0, \min(t_r, \tau_A))} \right\},
\end{aligned}$$

which is the same as the case when $0 < t_r < \tau_F < t_s \leq t_R$ in 2). Additionally from (B.17) we have

$$\pi(t_r) = P(\Delta_i(t_r) = 1) = P(l_i \leq t_r \mid l_i \leq \tau_A)$$

$$\begin{aligned}
&= \frac{\int_0^{t_r} \lambda (\lambda l \phi + 1)^{-\phi^{-1}-1} dl}{\int_0^{\tau_A} \lambda (\lambda l \phi + 1)^{-\phi^{-1}-1} dl} = \frac{\lambda \left(\frac{1}{-\lambda} (\lambda l \phi + 1)^{-\phi^{-1}} \right) \Big|_0^{t_r}}{\lambda \left(\frac{1}{-\lambda} (\lambda l \phi + 1)^{-\phi^{-1}} \right) \Big|_0^{\tau_A}} = \frac{1 - (\lambda \phi t_r + 1)^{-\phi^{-1}}}{1 - (\lambda \phi \tau_A + 1)^{-\phi^{-1}}}.
\end{aligned}$$

B.7 Supplementary Table of One-sided Boundaries

Table B.1 presents the one-sided boundaries for a non-inferiority trial design of a plasma donation study as discussed in Section 3.11.2. The parameter settings are as follows, $\lambda = 21$, $\phi = 0.8$, $\mu_0 = 0.035\%$, $\exp(\beta_{1A}) = 1.5$, $\exp(\beta_{1M}) = 2.5$, $\rho = 0.005$, $\tau_A = 0.5$,

$\tau_F = 0.5$, $\alpha = 5\%$, and target power=80%. Boundaries based on O'Brien & Fleming, Pocock, and uniform-type alpha spending functions are listed. Two interim analyses with one final analysis are planned; 1st interim analysis is planned at study time t_1 , 2nd interim analysis is planned at study time t_2 , and the final analysis is planned at the end of the study. The information fraction is determined as the fraction of inverse variance of $\hat{\beta}_1$.

Table B.1: One-sided boundaries from different alpha spending functions for a non-inferiority trial with 1st interim analysis planned at study time t_1 , 2nd interim analysis planned at study time t_2 , and final analysis planned at the end of study; information fraction is determined as the fraction of inverse variance of $\hat{\beta}_1$; $\lambda = 21$, $\phi = 0.8$, $\mu_0 = 0.035\%$, $\exp(\beta_{1A}) = 1.5$, $\exp(\beta_{1M}) = 2.5$, $\rho = 0.005$, $\tau_A = 0.5$, $\tau_F = 0.5$, $\alpha = 5\%$, and target power=80%.

(a) O'Brien & Fleming Type ($\alpha_1^*(t)$)										
1st interim analysis at t_1	2nd interim analysis at t_2									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
	(5.08, -, 1.64)	(5.08, 3.25, 1.65)	(5.08, 2.51, 1.66)	(5.08, 2.09, 1.70)	(5.08, 1.81, 1.75)	(5.08, 1.70, 1.78)	(5.08, 1.67, 1.84)	(5.08, 1.65, 1.96)	(5.08, 1.65, 2.15)	(5.08, 1.64, -)
	-	(3.25, -, 1.65)	(3.25, 2.52, 1.66)	(3.25, 2.09, 1.70)	(3.25, 1.81, 1.75)	(3.25, 1.70, 1.78)	(3.25, 1.67, 1.84)	(3.25, 1.65, 1.96)	(3.25, 1.65, 2.15)	(3.25, 1.65, -)
	-	-	(2.51, -, 1.66)	(2.51, 2.13, 1.70)	(2.51, 1.83, 1.76)	(2.51, 1.72, 1.79)	(2.51, 1.69, 1.85)	(2.51, 1.67, 1.97)	(2.51, 1.66, 2.16)	(2.51, 1.66, -)
	-	-	-	(2.09, -, 1.70)	(2.09, 1.88, 1.77)	(2.09, 1.76, 1.81)	(2.09, 1.72, 1.87)	(2.09, 1.71, 1.99)	(2.09, 1.70, 2.18)	(2.09, 1.70, -)
	-	-	-	-	(1.81, -, 1.75)	(1.81, 1.81, 1.82)	(1.81, 1.77, 1.88)	(1.81, 1.76, 1.99)	(1.81, 1.75, 2.17)	(1.81, 1.75, -)
	-	-	-	-	-	(1.70, -, 1.78)	(1.70, 1.81, 1.85)	(1.70, 1.79, 1.94)	(1.70, 1.78, 2.09)	(1.70, 1.78, -)
	-	-	-	-	-	-	(1.67, -, 1.84)	(1.67, 1.86, 1.92)	(1.67, 1.84, 2.06)	(1.67, 1.84, -)
	-	-	-	-	-	-	-	(1.65, -, 1.96)	(1.65, 1.97, 2.08)	(1.65, 1.96, -)
	-	-	-	-	-	-	-	-	(1.65, -, 2.15)	(1.65, 2.15, -)
	-	-	-	-	-	-	-	-	-	(1.64, -, -)
(b) Pocock Type ($\alpha_2^*(t)$)										
1st interim analysis at t_1	2nd interim analysis at t_2									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
	(2.29, -, 1.73)	(2.29, 2.17, 1.83)	(2.29, 1.98, 1.91)	(2.29, 1.86, 1.96)	(2.29, 1.78, 1.99)	(2.29, 1.75, 1.97)	(2.29, 1.74, 2.02)	(2.29, 1.73, 2.14)	(2.29, 1.73, 2.33)	(2.29, 1.73, -)
	-	(2.01, -, 1.82)	(2.01, 2.09, 1.91)	(2.01, 1.96, 1.98)	(2.01, 1.87, 2.03)	(2.01, 1.83, 2.03)	(2.01, 1.82, 2.08)	(2.01, 1.82, 2.20)	(2.01, 1.82, 2.39)	(2.01, 1.82, -)
	-	-	(1.86, -, 1.88)	(1.86, 2.03, 1.98)	(1.86, 1.94, 2.05)	(1.86, 1.90, 2.07)	(1.86, 1.89, 2.12)	(1.86, 1.89, 2.24)	(1.86, 1.88, 2.42)	(1.86, 1.88, -)
	-	-	-	(1.76, -, 1.93)	(1.76, 1.97, 2.03)	(1.76, 1.95, 2.08)	(1.76, 1.94, 2.14)	(1.76, 1.93, 2.25)	(1.76, 1.93, 2.43)	(1.76, 1.93, -)
	-	-	-	-	(1.69, -, 1.93)	(1.69, 1.93, 2.02)	(1.69, 1.94, 2.08)	(1.69, 1.93, 2.18)	(1.69, 1.93, 2.35)	(1.69, 1.93, -)
	-	-	-	-	-	(1.66, -, 1.91)	(1.66, 1.90, 1.98)	(1.66, 1.91, 2.06)	(1.66, 1.91, 2.20)	(1.66, 1.91, -)
	-	-	-	-	-	-	(1.65, -, 1.95)	(1.65, 1.96, 2.03)	(1.65, 1.96, 2.15)	(1.65, 1.95, -)
	-	-	-	-	-	-	-	(1.65, -, 2.07)	(1.65, 2.08, 2.18)	(1.65, 2.07, -)
	-	-	-	-	-	-	-	-	(1.65, -, 2.26)	(1.65, 2.26, -)
	-	-	-	-	-	-	-	-	-	(1.64, -, -)
(c) Uniform Type ($\alpha_3^*(t)$)										
1st interim analysis at t_1	2nd interim analysis at t_2									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
	(2.45, -, 1.70)	(2.45, 2.27, 1.76)	(2.45, 2.04, 1.82)	(2.45, 1.89, 1.87)	(2.45, 1.77, 1.90)	(2.45, 1.72, 1.90)	(2.45, 1.71, 1.96)	(2.45, 1.70, 2.07)	(2.45, 1.70, 2.26)	(2.45, 1.70, -)
	-	(2.14, -, 1.75)	(2.14, 2.14, 1.82)	(2.14, 1.96, 1.88)	(2.14, 1.83, 1.93)	(2.14, 1.78, 1.94)	(2.14, 1.77, 2.00)	(2.14, 1.76, 2.11)	(2.14, 1.76, 2.30)	(2.14, 1.75, -)
	-	-	(1.95, -, 1.81)	(1.95, 2.03, 1.88)	(1.95, 1.89, 1.95)	(1.95, 1.84, 1.97)	(1.95, 1.82, 2.03)	(1.95, 1.81, 2.14)	(1.95, 1.81, 2.33)	(1.95, 1.81, -)
	-	-	-	(1.82, -, 1.85)	(1.82, 1.93, 1.94)	(1.82, 1.88, 1.98)	(1.82, 1.86, 2.04)	(1.82, 1.85, 2.15)	(1.82, 1.85, 2.34)	(1.82, 1.85, -)
	-	-	-	-	(1.71, -, 1.87)	(1.71, 1.89, 1.95)	(1.71, 1.88, 2.01)	(1.71, 1.87, 2.11)	(1.71, 1.87, 2.28)	(1.71, 1.87, -)
	-	-	-	-	-	(1.67, -, 1.87)	(1.67, 1.86, 1.93)	(1.67, 1.87, 2.02)	(1.67, 1.87, 2.16)	(1.67, 1.87, -)
	-	-	-	-	-	-	(1.65, -, 1.92)	(1.65, 1.93, 2.00)	(1.65, 1.92, 2.12)	(1.65, 1.92, -)
	-	-	-	-	-	-	-	(1.65, -, 2.03)	(1.65, 2.04, 2.15)	(1.65, 2.03, -)
	-	-	-	-	-	-	-	-	(1.65, -, 2.22)	(1.65, 2.22, -)
	-	-	-	-	-	-	-	-	-	(1.64, -, -)

Appendix C

Appendix to Chapter 4

C.1 Derivations of Score Function

For simplicity of notation, we omit the subscript i for the derivation of (4.4),

$$\begin{aligned} \mathbf{S}_1(\boldsymbol{\vartheta}) &= \frac{\partial \ell(\boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}'} = R\mathcal{S}_1(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) + (1 - R) \frac{\frac{\partial}{\partial \boldsymbol{\theta}} \int P(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) P(X_1 \mid \mathbf{X}_2; \boldsymbol{\gamma}_1) dx_1}{\int P(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) P(X_1 \mid \mathbf{X}_2; \boldsymbol{\gamma}_1) dx_1} \\ &= R\mathcal{S}_1(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) + (1 - R) \frac{\int \frac{\partial}{\partial \boldsymbol{\theta}} P(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) P(X_1 \mid \mathbf{X}_2; \boldsymbol{\gamma}_1) dx_1}{P(\mathcal{D} \mid \mathbf{X}_2; \boldsymbol{\theta}, \boldsymbol{\gamma}_1)}. \end{aligned}$$

As

$$\partial P(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) / \partial \boldsymbol{\theta} = P(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) \frac{\partial \log P(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

we have

$$\begin{aligned} \mathbf{S}_1(\boldsymbol{\vartheta}) &= \frac{\partial \ell(\boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}'} = R\mathcal{S}_1(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) + (1 - R) \frac{\int \mathcal{S}_1(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) P(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) P(X_1 \mid \mathbf{X}_2; \boldsymbol{\gamma}_1) dx_1}{P(\mathcal{D} \mid \mathbf{X}_2; \boldsymbol{\theta}, \boldsymbol{\gamma}_1)} \\ &= R\mathcal{S}_1(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) + (1 - R) \frac{\int \mathcal{S}_1(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) P(\mathcal{D}, X_1 \mid \mathbf{X}_2; \boldsymbol{\theta}, \boldsymbol{\gamma}_1) dx_1}{P(\mathcal{D} \mid \mathbf{X}_2; \boldsymbol{\theta}, \boldsymbol{\gamma}_1)} \\ &= R\mathcal{S}_1(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) + (1 - R) \int \mathcal{S}_1(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) P(X_1 \mid \mathbf{X}_2, \mathcal{D}; \boldsymbol{\theta}, \boldsymbol{\gamma}_1) dx_1 \\ &= R\mathcal{S}_1(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) + (1 - R) E_{X_1} \{ \mathcal{S}_1(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) \mid \mathcal{D}, \mathbf{X}_2; \boldsymbol{\theta}, \boldsymbol{\gamma}_1 \}, \end{aligned}$$

where

$$S_1(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) = \frac{\partial \log P(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} .$$

$S_2(\boldsymbol{\vartheta})$ can be derived similarly.

C.2 Derivations of Score-Type Residuals M_μ

Let $\mu = X_1\beta_1 + \mathbf{X}'_2\boldsymbol{\beta}_2$. We have

$$\begin{aligned} \frac{\partial P_{klr}(\mathbf{X})}{\partial \mu} &= \sum_{j=k}^l \frac{\partial C_{kjl} e^{-\lambda_j(a_r - a_{r-1})}}{\partial \mu} \\ &= \sum_{j=k}^l \frac{\partial}{\partial \mu} \frac{e^{\mu(l-k)} \prod_{s=k}^{l-1} \lambda_{s0}}{e^{\mu(l-k)} \prod_{s=k, s \neq j}^l (\lambda_{s0} - \lambda_{j0})} e^{-\lambda_j(a_r - a_{r-1})} \\ &= \sum_{j=k}^l \frac{\prod_{s=k}^{l-1} \lambda_{s0}}{\prod_{s=k, s \neq j}^l (\lambda_{s0} - \lambda_{j0})} \frac{\partial}{\partial \mu} e^{-\lambda_j(a_r - a_{r-1})} \\ &= \sum_{j=k}^l \frac{\prod_{s=k}^{l-1} \lambda_{s0}}{\prod_{s=k, s \neq j}^l (\lambda_{s0} - \lambda_{j0})} e^{-\lambda_j(a_r - a_{r-1})} \{-\lambda_j(a_r - a_{r-1})\} \\ &= \sum_{j=k}^l P_{klr}^{(j)}(\mathbf{X}) \{-\lambda_j(a_r - a_{r-1})\} , \quad k \leq l \end{aligned}$$

where

$$\begin{aligned} P_{klr}^{(j)}(\mathbf{X}) &= C_{kjl} e^{-\lambda_j(a_r - a_{r-1})} , \\ P_{klr}(\mathbf{X}) &= \sum_{j=k}^l P_{klr}^{(j)}(\mathbf{X}) . \end{aligned}$$

Hence we have

$$\begin{aligned}
\log P(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta}) &= \sum_{r=1}^m \sum_{k=1}^K \sum_{l=1}^K \delta_{klr} \log P_{klr}(\mathbf{X}) , \\
M_{\mu} &= \frac{\partial \log P(\mathcal{D} \mid \mathbf{X}; \boldsymbol{\theta})}{\partial \mu} = \sum_{r=1}^m \sum_{k=1}^K \sum_{l=1}^K \frac{\delta_{klr}}{P_{klr}(\mathbf{X})} \frac{\partial P_{klr}(\mathbf{X})}{\partial \mu} \\
&= \sum_{r=1}^m \sum_{k=1}^K \sum_{l=1}^K \frac{\delta_{klr}}{P_{klr}(\mathbf{X})} \sum_{j=k}^l P_{klr}^{(j)}(\mathbf{X}) \{-\lambda_j(a_r - a_{r-1})\} \\
&= \sum_{r=1}^m \sum_{k=1}^K \sum_{l=1}^K \delta_{klr} \frac{\sum_{j=k}^l P_{klr}^{(j)}(\mathbf{X}) \{-\lambda_j(a_r - a_{r-1})\}}{P_{klr}(\mathbf{X})} \\
&= \sum_{r=1}^m \sum_{k=1}^K \sum_{l=1}^K -\delta_{klr}(a_r - a_{r-1}) \frac{\sum_{j=k}^l \lambda_j P_{klr}^{(j)}(\mathbf{X})}{P_{klr}(\mathbf{X})} , \quad k \leq l .
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E} \left\{ \sum_{l=1}^K \frac{I(Z(a_{r-1}) = k) I(Z(a_r) = l)}{P_{klr}(\mathbf{X})} \frac{\partial P_{klr}(\mathbf{X})}{\partial \mu} \middle| I(Z(a_{r-1}) = k) = 1 \right\} \\
&= \sum_{l=1}^K \mathbb{E} \left\{ \frac{I(Z(a_r) = l)}{P_{klr}(\mathbf{X})} \frac{\partial P_{klr}(\mathbf{X})}{\partial \mu} \middle| I(Z(a_{r-1}) = k) = 1 \right\} \\
&= \sum_{l=1}^K \frac{\partial P_{klr}(\mathbf{X})}{\partial \mu} I(Z(a_{r-1}) = k) \\
&= I(Z(a_{r-1}) = k) \frac{\partial}{\partial \mu} \left\{ \sum_{l=1}^K P_{klr}(\mathbf{X}) \right\} \\
&= 0 \text{ since } \sum_{l=1}^K P_{klr}(\mathbf{X}) = 1 ,
\end{aligned}$$

which gives

$$\mathbb{E} \left\{ \sum_{l=1}^K \frac{I(Z(a_{r-1}) = k)I(Z(a_r) = l)}{P_{klr}(\mathbf{X})} \frac{\partial P_{klr}(\mathbf{X})}{\partial \mu} \right\} = 0.$$

Hence the expectation of M_μ is

$$\begin{aligned} \mathbb{E}(M_\mu) &= \mathbb{E} \left\{ \sum_{r=1}^m \sum_{k=1}^K \sum_{l=1}^K \frac{\delta_{klr}}{P_{klr}(\mathbf{X})} \frac{\partial P_{klr}(\mathbf{X})}{\partial \mu} \right\} \\ &= \sum_{r=1}^m \sum_{k=1}^K \mathbb{E} \left\{ \sum_{l=1}^K \frac{\delta_{klr}}{P_{klr}(\mathbf{X})} \frac{\partial P_{klr}(\mathbf{X})}{\partial \mu} \right\} \\ &= \sum_{r=1}^m \sum_{k=1}^K \mathbb{E} \left\{ \sum_{l=1}^K \frac{I(Z(a_{r-1}) = k)I(Z(a_r) = l)}{P_{klr}(\mathbf{X})} \frac{\partial P_{klr}(\mathbf{X})}{\partial \mu} \right\} \\ &= 0. \end{aligned}$$

C.3 Supplementary Table of Strata Sizes

Table C.1 presents the average number of strata sizes based on 1000 replicates in the empirical studies of two-phase designs in Section 4.5 where the strata are defined according to the state observed at the last visit (i.e., state $Z(\tau) = 0, 1$, or 2) and the binary covariate X_2 . Strata are defined specifically as follows.

- stratum 1: $Z(\tau) = 0, X_2 = 0$
- stratum 2: $Z(\tau) = 0, X_2 = 1$
- stratum 3: $Z(\tau) = 1, X_2 = 0$
- stratum 4: $Z(\tau) = 1, X_2 = 1$
- stratum 5: $Z(\tau) = 2, X_2 = 0$
- stratum 6: $Z(\tau) = 2, X_2 = 1$

Table C.1: Average strata sizes based on 1000 replicates[†].

n_2	$P(Z(\tau) = 2)$	$P(X_1 = 1)$	Phase I								SRS						RES						BAL					
			β_1	s1	s2	s3	s4	s5	s6	s1	s2	s3	s4	s5	s6	s1	s2	s3	s4	s5	s6	s1	s2	s3	s4	s5	s6	
300	0.4	0.2	0	512	275	320	308	167	419	77	41	48	46	25	63	75	75	0	0	75	75	50	50	50	50	50	50	
			0.22	517	273	318	304	164	424	78	41	47	46	24	64	75	75	0	0	75	75	50	50	50	50	50	50	
		0.5	0.69	529	276	308	292	162	434	80	41	46	44	25	65	75	75	0	0	75	75	50	50	50	50	50	50	
			0	512	275	320	307	167	419	77	41	48	46	25	63	75	75	0	0	75	75	50	50	50	50	50	50	
	0.8	0.2	0.22	519	271	318	304	162	426	78	40	48	46	24	64	75	75	0	0	75	75	50	50	50	50	50	50	
			0.69	539	276	300	287	160	438	81	42	45	43	24	66	75	75	0	0	75	75	50	50	50	50	50	50	
		0.5	0.22	215	64	276	114	508	823	32	10	42	17	76	123	68	46	7	29	75	75	50	50	50	50	50	50	
			0.69	215	61	268	106	516	834	33	9	40	16	77	125	66	45	8	30	75	75	50	50	50	50	50	50	
		0.5	0.22	218	62	276	109	505	831	33	9	42	16	76	124	67	45	8	30	75	75	50	50	50	50	50	50	
			0.69	220	59	258	100	522	842	33	9	39	15	79	126	65	45	10	30	75	75	50	50	50	50	50	50	
600	0.4	0.2	0	512	275	320	307	167	419	154	83	96	92	50	125	150	150	0	0	150	150	100	100	100	100	100		
			0.22	517	273	318	304	164	424	155	82	95	91	49	127	150	150	1	0	149	150	100	100	100	100	100		
		0.5	0.69	529	276	308	292	162	434	159	83	92	88	49	130	150	150	1	0	149	150	100	100	100	100	100		
			0	512	275	320	307	167	419	153	83	96	92	50	125	150	150	0	0	150	150	100	100	100	100	100		
	0.8	0.2	0.22	519	271	318	304	162	426	156	81	95	91	49	128	150	150	1	0	149	150	100	100	100	100	100		
			0.69	538	276	300	287	160	438	162	83	90	86	48	131	150	150	1	0	149	150	100	100	100	100	100		
		0.5	0.22	215	64	276	114	508	823	64	19	83	34	152	247	122	64	28	84	150	152	108	64	108	106	107		
			0.69	215	61	268	106	516	834	65	18	80	32	154	250	122	61	28	85	150	153	109	61	109	103	109		
		0.5	0.22	218	62	276	114	508	823	64	19	83	34	152	247	123	64	27	84	150	152	108	64	108	106	107		
			0.69	220	59	258	100	522	842	66	18	78	30	157	252	122	59	28	85	150	156	111	59	111	99	110		

[†] Stratum 1: $Z(\tau) = 0, X_2 = 0$; stratum 2: $Z(\tau) = 0, X_2 = 1$; stratum 3: $Z(\tau) = 1, X_2 = 0$; stratum 4: $Z(\tau) = 1, X_2 = 1$; stratum 5: $Z(\tau) = 2, X_2 = 0$; stratum 6: $Z(\tau) = 2, X_2 = 1$.