On the Correction for Misclassification Bias in Electronic Health Data Using Validation Sample Approaches

by

Christopher A. Gravel

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Mathematics

School of Mathematics and Statistics
Ottawa-Carleton Institute for Mathematics and Statistics
CARLETON UNIVERSITY
April, 2015

Acknowledgments

To begin, I would like to thank both my co-supervisors for all their insight and help throughout the development of this work. Dr. Patrick Farrell and Dr. Daniel Krewski have guided me through this process and their advice and been invaluable. I have learned so much from both of them in ways that have propelled my evolution as a statistician and I will be forever grateful for all the support and encouragement they provided throughout my doctoral studies.

I would also like to thank Dr. Anup Dewanji at the Indian Statistical Institute, Kolkata. He provided much of the motivation for this thesis and assisted me in formulating my preliminary investigations into this work.

Next, I would like to extend my gratitude to the School of Mathematics and Statistics at Carleton University and to the faculty members and staff that have taught me so much. Particularly, Dr. Shirley Mills and Dr. Mohamedou Ould Haye for so much insight and inspiration and Cate Palmer for her assistance and support throughout my PhD. I learned a love for probability and statistics through the many courses and interactions with the faculty and I am extremely grateful.

I would like to thank my examining committee, Dr. Chul Guy Park, Dr. Paul Villeneuve, Dr. Tim Ramsay and Dr. Celia Greenwood, for taking the time to review my work and for providing a number of valuable suggestions and insights.

I want to thank my friends for helping me through stressful times and in particular my partner, Emilie Gravel, who kept me going even in some of the most difficult and stressful parts of my PhD. Her ability to empathize with the challenges I was facing and her love and support helped me to stay focused and strong.

Finally, I want to thank my parents, Dr. Roy Gravel and Ying Gravel, for their encouragement. They have seen me evolve, both personally and professionally,

over the many years I have been working on post-secondary education, and have given so much help in facilitating these changes. I am extremely grateful for all the love and support they have given me.

Abstract

Post-market drug safety researchers use large data sets comprised of individual patient electronic health records to assess potential adverse drug reaction risk. These records are assumed to have perfect classification of the outcomes of interest, however, this assumption is not necessarily realistic. There are a number of reasons for outcome misclassification to be present in electronic health records data. Coding issues, diagnostic uncertainty (particularly relative to time), and misdiagnoses are all possible causes of this form of measurement error.

Unbiased estimation with the presence of outcome misclassification relies on the availability of additional information. We considered the use of internally validated data for this purpose and demonstrated misclassification bias adjustment in binary data and right censored continuous time survival data with and without the presence of competing risks. These data structures are investigated as they pertain to the underlying nature of electronic health records datasets which is the motivating example for this research.

In misclassified binary data we considered the use of different sampling schemes for acquisition of the validation data. We first considered the estimated asymptotic relative efficiencies between the maximum likelihood estimators derived from these sampling approaches. Monte Carlo simulation demonstrated that the possibility of a minimal variance MLE relative to differing sampling schemes exists, however, the ability to assess this prior to sampling is not possible. Hence, we propose a numerical method that results in a validation sample size determination algorithm that can be used to approximate the relationship between sample size, variance of the estimator of the parameter of interest and the chosen sampling approach. Finally, we considered methods of estimation used to assess association in a two-by-two contingency table such as the odds-ratio and logistic regression.

For right censored continuous time survival data with and without competing risks, we propose the use of internal validation to adjust for misclassification bias of two different types. First, we considered the problem of failing to observe the occurrence of an event of interest and incorrectly concluding that the individual under study is a censored observation. Second, we considered the situation in which we correctly observe an event occurrence, however erroneously observe the cause-specific event type. Under assumptions based on the motivating example, using a multi-sample likelihood based approach we produced unbiased estimators for data with either form of error or both simultaneously being present.

Preface

This thesis is comprised of four manuscripts that make up my doctoral research project. These manuscripts are currently unpublished but are in preparation for submission. Each article has the same authorship, myself and my co-supervisors Dr. Patrick Farrell and Dr. Dan Krewski. The work presented in this thesis represents my own independent research, however, I have greatly benefited from the guidance and assistance of my co-supervisors as well as Dr. Anup Dewanji of the Indian Statistical Institute, Kolkata.

- Gravel, C.A., Farrell, P.J., Krewski, D. On the Optimization of a Validation Sampling Approach for Unbiased Estimation in Binary Data in the Presence of Outcome Misclassification. (In preparation).

- Gravel, C.A., Farrell, P.J., Krewski, D. Monte Carlo Sample Size Determination for Unbiased Estimation with Validation Data in the Presence of Binary Outcome Misclassification. (In preparation).

- Gravel, C.A., Farrell, P.J., Krewski, D. On the Optimization of a Validation Sampling Approach for Misclassification Bias Adjustment in Logistic Regression Models. (In preparation).

- Gravel, C.A., Farrell, P.J., Krewski, D. A Validation Sampling Approach for Unbiased Estimation in Misclassified Right Censored Continuous Time Survival Data With and Without Competing Risks. (In preparation).

The manuscripts are presented in a logical order and while some redundancy is unavoidable, they have been altered to avoid some repetition. Each chapter has its own introduction and conclusion written for flow within this thesis. A general literature review for all the articles is presented in Chapter 1 and the appendices and bibliographies are merged into single sections for this thesis.

Due to the large amount of results presented in Appendix D of this thesis, we shall provide it as a secondary document available upon request from the author (christophergravel@cmail.carleton.ca). The results contained in this additional document are similar to those presented in the individual chapters of the thesis, hence, for brevity we chose to exclude the majority of these tables.

Contents

Acknowledgments I

Abstract III

Preface V

List of Tables X

List of Figures XV

1 Introduction 1

1.1 Summary 1

1.2 Literature Review 7

1.3 Thesis Objectives 12

2 On the Optimization of a Validation Sampling Approach for Unbiased Estimation in Binary Data in the Presence of Outcome Misclassification 14

2.1 Introduction 14

2.2 Binomial Likelihood with Outcome Misclassification 16

2.3 A Multi-Sample Framework for MCB-Adjusted Estimation 21

2.3.1 Comparative Simulation Studies: 2S versus 3S Approaches 24

2.4 The Inverse Variance Weighted Estimator of the Odds-Ratio 32

2.5 Comparative Simulation Study: IVW, 2S and 3S Approaches 35

2.6 Conclusions and Discussion 49

3 Monte Carlo Sample Size Determination for Unbiased Estimation with Validation Data in the Presence of Binary Outcome Misclassification . 51

3.1 Introduction 51

3.2 Asymptotic Properties for Multi-Sample Maximum Likelihood Estimation . . . 53
3.3 Development of a Monte Carlo Sample Size Determination Algorithm . 54
3.4 Simulation Study to Investigate Monte Carlo Sample Size Determination Algorithm . . . 65
3.5 Conclusions and Discussion . . . 72
4 On the Optimization of a Validation Sampling Approach for Misclassification Bias Adjustment in Logistic Regression Models . . . 74
4.1 Introduction . . . 74
4.2 Binomial Likelihood with Outcome Misclassification and Logistic Regression . . . 76
4.3 Comparative Simulation Study for Logistic Regression: 2S versus 3S Approach . . . 82
4.4 Monte Carlo Sample Size Determination Algorithm with Extensions for Logistic Regression . . . 98
4.5 Conclusions and Discussion . . . 106
5 A Validation Sampling Approach for Unbiased Estimation in Misclassified Right Censored Continuous Time Survival Data With and Without Competing Risks . . . 108
5.1 Introduction . . . 108
5.2 Notation and Likelihood Construction . . . 111
5.3 A Univariate Model, $L^{(1)}$ . . . 113
5.4 A Competing Risks Model, $L^{(2)}$ . . . 114
5.5 Asymptotic Properties for Multi-Sample Maximum Likelihood Estimation . . . 119
5.6 Simulation Studies to Investigate the Properties of $L^{(1)}$ and $L^{(2)}$ . . . 122
5.7 Results for Simulation Studies Investigating $L^{(1)}$ . . . 126
5.8 Results for Simulation Studies Investigating $L^{(2)}$ . . . 133
5.9 Conclusions and Discussion . . . 142

A Mathematical Results . . . 145
A.1 Preliminary Results . . . 145
A.2 Multi-Sample Results . . . 147
B Derivation of Results for Chapter 5 . . . 155
B.1 Univariate Model, $L^{(1)}$ . . . 155
B.2 Competing Risks Model, $L^{(2)}$ . . . 158
B.3 Competing Risks Model, $L^{(2b)}$ . . . 159
C R code . . . 160
C.1 Chapter 2 Code . . . 160
C.2 Chapter 3 Code . . . 170
C.3 Chapter 4 Code . . . 172
C.4 Chapter 5 Code . . . 192
D Additional Tables . . . 210
Bibliography . . . 211

List of Tables

2.2.1 Cell probabilities in the 2x2 cross-classification of the source population. 16

2.2.2 Cell probabilities in the 2x2 cross-classification of the observed data. . 17

2.2.3 Summary table of cell counts for the original and validation data. . . 18

2.3.1 Chosen target parameters for the simulation studies comparing the 2S and 3S approaches. 26

2.3.2 Simulation study results for the 2S versus 3S approach for parameter set 2. 28

2.3.3 Simulation study results for the 2S versus 3S approach for parameter set 3. 28

2.5.1 Comparative simulation study results of odds-ratios estimated using 2S, 3S and IVW for parameter set 1 (OR = 0.4737). 35

2.5.2 Comparative simulation study results of odds-ratios estimated using 2S, 3S and IVW for parameter set 2 (OR = 0.4444). 35

2.5.3 Comparative simulation study results of odds-ratios estimated using 2S, 3S and IVW for parameter set 3 (OR = 1.7143). 36

2.5.4 Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 1$, $n = 10,000$, $m = 1,000$. 39

2.5.5 Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 0.5$, $n = 10,000$, $m = 1,000$. 39

2.5.6 Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 2$, $n = 10,000$, $m = 1,000$. 40

2.5.7 Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for increasing values of $\frac{n_{.1}}{n}$, $n = 10,000$, $m = 1,000$. 40

2.5.8 Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 1$, $n = 10,000$, $m = 1,000$. 40

2.5.9 Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 0.5$, $n = 10,000$, $m = 1,000$. 41

2.5.10 Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 2$, $n = 10,000$, $m = 1,000$. 41

2.5.11 Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for incidence rates greater than 0.5, $OR = 2$, $n = 10,000$, $m = 1,000$. 42

2.5.12 Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for incidence rates greater than 0.5, $OR = 2$, $n = 10,000$, $m = 1,000$. 42

2.5.13 Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 2$, $n = 100,000$, $m = 1,000$. 43

2.5.14 Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 2$, $n = 100,000$, $m = 1,000$. 43

2.5.15 Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 1$, $n_{1.} = 8000$, $n_{0.} = 2000$, $m = 1,000$. . 46

2.5.16 Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 1$, $n_{1.} = 2000$, $n_{0.} = 8000$, $m = 1,000$. . 47

3.3.1 Chosen target parameters for the simulation studies comparing the 2S and 3S approaches. 58

3.4.1 Simulation study results for the 2S versus 3S approach for parameter set 1. 65

3.4.2 Simulation study results for the 2S versus 3S approach for parameter set 2. 65

3.4.3 Simulation study results for the 2S versus 3S approach for parameter set 3. 65

3.4.4 Monte Carlo sample size determination for parameter set 1 with n= 10,000. 66

3.4.5 Monte Carlo sample size determination for parameter set 2 with n= 10,000. 67

3.4.6 Monte Carlo sample size determination for parameter set 3 with n= 10,000. 68

4.3.1 Chosen target π-parameters for the simulation study comparing the 2S and 3S approach for logistic regression. 86

4.3.2 Chosen target θ and τ-parameters for the simulation study comparing the 2S and 3S approach for logistic regression. 87

4.3.3 Target regression parameters for the simulation study comparing the 2S and 3S approach for logistic regression. 87

4.3.4 Results for parameter set 1 comparing the 2S and 3S approach for logistic regression with n=10,000, m=2,000. 88

4.3.5 Results for parameter set 1 comparing the 2S and 3S approach for logistic regression with n=10,000, m=3,000. 89

4.3.6 Results for parameter set 3 comparing the 2S and 3S approach for logistic regression with n=10,000, m=2,000. 90

4.3.7 Results for parameter set 3 comparing the 2S and 3S approach for logistic regression with n=10,000, m=3,000. 91

4.3.8 Results for parameter set 3 comparing the 2S and 3S approach for logistic regression with n=10,000, m=5,000. 92

4.3.9 Chosen target regression parameters for additional simulation study comparing the 2S and 3S approach for logistic regression. 96

4.4.1 Monte Carlo sample size determination for parameter set 1 from Table 4.3.3 with n = 10,000. 100

4.4.2 Monte Carlo sample size determination for parameter set 3 from Table 4.3.3 with n = 10,000. 101

4.4.3 Expected subgroup sample sizes for parameter set 3 from Table 4.3.3. 104

5.7.1 Target parameter for simulation study investigating univariate model, $L^{(1)}$. 126

5.7.2 Results for the simulation study investigating univariate model, $L^{(1)}$, for parameter set 1 with n = 10000. 127

5.7.3 Results for the simulation study investigating univariate model, $L^{(1)}$, for parameter set 2 with n = 10000. 128

5.7.4 Results for the simulation study investigating univariate model, $L^{(1)}$, for parameter set 3 with n = 10000. 129

5.7.5 Results for the simulation study investigating univariate model, $L^{(1)}$, for parameter set 4 with n = 10000. 130

5.7.6 Results for the simulation study investigating univariate model, $L^{(1)}$, for parameter set 5 with n = 10000. 131

5.7.7 Results for the simulation study investigating univariate model, $L^{(1)}$, for parameter set 4 with n = 1000, m = 100. 131

5.7.8 Results for the simulation study investigating univariate model, $L^{(1)}$, for parameter set 4 with n = 5000, m = 500. 132

5.8.1 Target parameters for simulation study investigating competing risks model, $L^{(2)}$. 134

5.8.2 Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 1 with n = 10000. 135

5.8.3 Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 2 with n = 10000. 136

5.8.4 Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 3 with n = 10000. 137

5.8.5 Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 4 with n = 10000. 138

5.8.6 Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 5 with n = 10000. 139

5.8.7 Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 5 with n = 1000, m = 200. 140

5.8.8 Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 5 with n = 5000, m = 1000. 140

5.8.9 Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 4 with n = 10000, m1 = 300, m0 = 700. . . . 140

5.8.10 Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 4 with n = 10000, m1 = 700, m0 = 300. . . . 141

List of Figures

2.3.1 Plot of Monte Carlo estimates of $\max(\hat{\sigma}(\hat{\pi}_{1|1}), \hat{\sigma}(\hat{\pi}_{1|0}))$ versus P_1 for parameter set 1. 30

2.3.2 Plot of Monte Carlo estimates of $\max(\hat{\sigma}(\hat{\pi}_{1|1}), \hat{\sigma}(\hat{\pi}_{1|0}))$ versus P_1 for parameter set 3. 31

2.5.1 Plot of Monte Carlo estimates of $\max(\hat{\sigma}(\hat{\pi}_{1|1}), \hat{\sigma}(\hat{\pi}_{1|0}))$ versus P_1 for example 1 (third row of Table 2.5.14). 45

2.5.2 Plot of Monte Carlo estimates of $\max(\hat{\sigma}(\hat{\pi}_{1|1}), \hat{\sigma}(\hat{\pi}_{1|0}))$ versus P_1 for example 2 (fourth row of Table 2.5.14). 45

3.3.1 Convergence behaviour of the validation subgroup's observed information for parameter set 1. 59

3.3.2 Convergence behaviour of the validation subgroup's observed information for parameter set 2. 59

3.3.3 Convergence behaviour of the validation subgroup's observed information for parameter set 3. 60

3.3.4 Convergence behaviour of observed information due to validation sampling differences for parameter set 1. 61

3.3.5 Convergence behaviour of observed information due to validation sampling differences for parameter set 2. 61

3.3.6 Convergence behaviour of observed information due to validation sampling differences for parameter set 3. 62

3.4.1 Plot of validation sample size relative to 2S and 3S sampling approaches for parameter set 1. Two bounds are considered, $c = 0.05, 0.1$ with an expected incidence rate of 0.097. 69

3.4.2 Plot of validation sample size relative to 2S and 3S sampling approaches for parameter set 2. Two bounds are considered, $c = 0.05, 0.1$ with an expected incidence rate of 0.195. 70

3.4.3 Plot of validation sample size relative to 2S and 3S sampling approaches for parameter set 3. Two bounds are considered, $c = 0.05, 0.1$ with an expected incidence rate of 0.3. 70

4.3.1 Plot of Monte Carlo Estimates of $\hat{\sigma}(\hat{\beta})$ versus P_1 for parameter sets 1 to 4. 94

4.3.2 Plot of Monte Carlo Estimates of $\hat{\sigma}(\hat{\beta})$ versus P_1 for parameter sets 5 to 8. 94

4.3.3 Plot of Monte Carlo Estimates of $\hat{\sigma}(\hat{\beta})$ versus P_1 for parameter sets 9 to 12. 95

4.3.4 Plot of Monte Carlo Estimates of $\hat{\sigma}(\hat{\beta})$ versus P_1 for additional target parameters from Table 4.3.9. 96

Chapter 1

Introduction

1.1 Summary

Numerous health organizations in a variety of countries routinely collect anonymized electronic health records (EHRs). An EHR is “a repository of patient [health] data in digital form, stored and exchanged securely, and accessible by multiple authorized users,” Häyrinen et al. (2008) [16]. These records are automatically gathered by health care professionals in hospitals and other health-related environments. They contain detailed patient data, coded using internationally developed terminologies, designed to classify the information describing encounters with the health care system. Diagnoses, lab testing, medication prescriptions, daily charting in hospitals, and a variety of other useful information make up the descriptive content in an EHR. This data source is longitudinal, where all patient encounters are assumed to be time stamped.

This source of patient information has enormous potential for health research, particularly in the area of pharmacovigilance (PhV) that focuses on the post-market detection of adverse effects associated with pharmaceutical products. Following the introduction of a new drug into the marketplace, there is a need for ongoing evaluation of the potential for adverse drug reactions (ADRs) under real world conditions of use that may not have been seen in limited clinical trials conducted prior to market authorization. Historically, information on the occurrence of adverse events (AEs) were derived from spontaneous reporting systems, wherein case reports of AEs were reported either indirectly through health care professionals, or directly by the individual experiencing the ADR. A number of inherent problems exist in this form of data collection; notably the voluntary nature of AE reporting, which violates random sampling assumptions. Further, Roux et al. (2005) [37] note that AE background incidence rates, the number of patients exposed to the drug, the extent of under-reporting and the true status of the drug/AE relationship are unknown when considering spontaneous reporting. Thus, it is desirable to investigate the relationship between drug utilization and the occurrence of ADRs using EHR datasets.

However, to adequately estimate this relationship, an unrealistic assumption of perfect coding and classification of the AE status in the EHRs must be met. Misclassification of the outcome variable can produce misclassification bias (MCB), which can shift the estimates of association or risk. MCB is a well documented problem in the literature and can occur for a number of reasons. Poor record keeping practices are a potential reason for the introduction of MCB. For example, Nicholson et al. (2011) [33] discuss variation in coding practices in the UK's General Practice Research Database as the data is either entered by professional coders interpreting clinical records or by the clinicians themselves as a part of patient "routine care". Another source of misclassification in the data is variation in coding definitions. For instance, diagnostic coding is done using internationally developed systems such as the 'International Classification of Diseases' (ICD) diagnostic tool developed by the world health organization (WHO) [40]. This coding system has undergone a number of revisions (the most recent being the tenth revision, denoted as ICD-10) and is described as extremely granular, in that coding an AE is done by breaking the event into many fine-detailed terms. In other words, a particular AE can be described in a variety of ways, and the presence of one particular code may not be sufficient. For example, Drahos et al. (2013) [6]

attempted to validate the ICD-9-CM coding system by estimating the diagnostic positive predictive value (PPV) of two infection-related conditions through comparison to a 'gold standard', which is assumed to contain the true outcomes not subject to misclassification error. They used two separate diagnostic definitions of these infections, one being the presence of two or more relevant codes, and the other being the presence of one code and a relevant prescription. Levine et al. (2013) [22] performed a similar study and utilized three different definitions of skin and soft tissue infections. Both of these studies demonstrate less-than-ideal estimates of diagnostic PPV, which will introduce a form of MCB to pharmacovigilance research, solely based upon variation between diagnostic definitions.

Misdiagnoses are another potential cause of misclassification of the outcome of interest in EHRs. These may be due to diagnostic errors arising from the use of error-prone diagnostic tests that are selected for use due to lower costs or minimal invasiveness, or errors made by the diagnostician in interpreting test results. However, another type of misclassification may be present in the data due to the fact that we are able to observe AEs as a function of time. For a given AE of interest, we can either treat the data as coming from a fixed time interval and observe the presence or absence of the outcome, or consider a time-to-event structure and estimate the time to first occurrence of the AE. Under both assumptions, we are waiting for a particular outcome, whose observation may require a wide array of testing. As such, misdiagnoses can occur for reasons other than diagnostic error. For example, consider the case in which an individual with symptoms presents him or herself at an emergency room, and the hospital decides to admit the patient. Then, during the time that the individual is under inpatient care, he or she is subjected to a variety of tests and interactions with diagnosticians. In this case it is possible that there will be an occurrence of misclassification in the EHR due to a difference between the time to a possible diagnosis, and the time to entry of the record or confirmation of that diagnosis. Imagine a diagnostic test that strongly suggests the occurrence of the AE of interest, but the final diagnosis is not recorded as further tests are ordered. Meanwhile, the period of observation ends, or perhaps the patient chooses to go to another health care facility and is lost to follow up. Another possibility is that an alternate AE is diagnosed, even though the AE of interest would be diagnosed had additional testing occurred. Here, a partial diagnosis has occurred during the period of observation and validation of the diagnoses would demonstrate misclassification in the EHRs.

Thus, there exists a variety of reasons for the presence of misclassification of the outcome of interest, AE occurrence, in EHR datasets. With this in mind, we wish to consider methods of adjusting for MCB in estimation of the ADR risk/association. Since the presence of misclassification may be unobservable in the EHR data, we require additional information that we can access through the use of an internal or external validation sample. Under this approach, an additional sample is collected using a measurement device that is assumed to be without error, which gives us access to the requisite information. An externally-drawn validation sample may provide estimates of the outcome misclassification rates in the source population. Whereas an internal validation sample will allow us to measure a subset of the original data, thereby providing an estimate of these rates from an information source housed in the original sample. Both methods require the assumption that the rates of misclassification are similar in the original data as well as the validation data. Since the internal validation data are a (randomly selected) subset of the original sample, this assumption seems reasonable in this case; however, it seems harder to justify that assertion with external validation data. Greenland (1988) [12] makes the claim, when discussing binary data, that internal validation is "clearly preferable to the use of external estimates". He notes that even when the misclassification rates are known in the source population, a binary classifier incorporating validation information may still perform poorly, further implying that the external information may not be representative of the rates in the original data. In the remainder of this discussion, we will assume that we have access to both the original EHR data as well as an internally-gathered validation sample in which outcomes are measured by an error-free device. The statistical challenge is then how to best design the validation sampling scheme.

Another aspect of outcome misclassification that needs to be considered is the notion of differential versus non-differential misclassification. Returning to our previous example, differential misclassification error would allow for the presence of different rates of misclassification to be possible in different drug utilization groups. This seems like a reasonable assumption in general. For instance, if a drug is erroneously suspected to have a serious adverse side effect, clinicians may be more likely to diagnose the AE. Since the assumption of non-differentiality introduces a constraint in estimation which may or may not be realistic we shall only consider differential misclassification in this discussion.

Using the validated data, this integrated thesis focuses on the development of techniques for adjusting for MCB with the objective of producing unbiased estimates of ADR risk/association. In Chapter 2, we extend previous work on the estimation of the odds-ratio, in the presence of misclassification, by introducing a validation sampling scheme that conditionally selects observations based on their original AE status, rather than selecting a random sample of all patients housed in the EHR data. This sampling approach allows the experimenter to control the categorical breakdown of the validation sample by specifying the number of observations falling into the two outcome categories of the original observed AE status. For this discussion, we will refer to the observations with the outcome status of interest (the individuals under study) as category 1 observations and the other group as category 2. For EHR data, this is an attractive approach. Diagnostic tests used to validate the original data can be extremely expensive there may be a large discrepancy between the cost of validating data for individuals that are originally observed to have experienced the AE versus those that have not. Thus, the ideal sample may be one that minimizes cost by sampling conditional on AE status, rather than using a simple random sample of all individuals, regardless of

AE status, which will tend to produce a categorical breakdown that is similar to that in the original data. Further, we will demonstrate that by carefully planning the validation sampling scheme it is possible to produce a better maximum likelihood estimator (MLE) with respect to asymptotic relative efficiency (ARE) with a conditional validation sample, as compared to a purely random validation sample. To demonstrate this, we outline the asymptotic properties of these estimators using a multi-sample framework (see Appendix A) to address the combination of information contained in two or more samples using a likelihood based approach. The statistical properties of the MLEs under different validation sampling schemes are then investigated in detail through Monte Carlo simulation.

In Chapter 3, we address the problem of sample size determination when using a conditional sampling approach to draw a validation sample. In Chapter 2, we demonstrate that the efficiency of the resulting estimators is dependent on the chosen categorical breakdown, a quantity that is specified prior to drawing a validation sample. Thus, we create an algorithm that uses a simulation-based approach to approximate the relationship between three quantities; the overall validation sample size, the number of category 1 observations chosen in the validation sample and a bound on a Bonferonni interval associated with the parameters of interest. This algorithm is designed specifically for contingency table approaches based on the estimation of the binomial success probabilities, and allows for the practical implementation of the algorithm for sample size determination in the context of the conditional validation sampling approach.

In Chapter 4, we extend the literature on logistic regression with the presence of outcome misclassification to allow for the application of the conditional sampling approach introduced in Chapters 2 and 3. We again consider comparative simulation studies and investigate the estimated AREs of the resulting MLEs under different validation sampling schemes. Although modelling the data within a multi-sample framework allows for a similar discussion of the asymptotic properties of the MLEs, the problem of sample size determination is harder to address in this context. Nevertheless, for simple models the methodology proposed in Chapter 3 is extended, and discussion surrounding the limitations of the resulting techniques and the lack of single general algorithm are addressed.

In Chapter 5, we introduce a novel validation sampling approach to adjust for MCB with right censored continuous time survival data, under certain assumptions regarding patterns of misclassification in EHR data. For this discussion, we use a parametric approach, first by only considering the time to the occurrence of a single AE of interest. We subsequently extend this to account for presence of competing risks and apply techniques for adjusting for MCB under a cause-specific (C-S) hazards formulation of competing risks with the presence of right censorship (Kalbfleish and Prentice (2002) [19]). This accounts for two types of misclassification; the first for the time to AE occurrence, the second for error in the observed C-S AE type. Once again, we model the use of additional validation information under a multi-sample framework and investigate the asymptotic properties. We then conduct simulation studies to numerically investigate these properties and the performance of these methods.

1.2 Literature Review

The problem of adjusting for differential binary misclassification using internal validation data has been investigated by many authors; however, the focus of this work has been primarily on exposure misclassification. In other words, the item subject to misclassification in our motivating PhV example would be drug utilization status, as opposed to the indicator of AE occurrence. However, we are interested in outcome misclassification and as such, we will first point out that all the results on binary exposure misclassification in the literature will hold for binary outcome misclassification. All variables are indicators and the probabilities of interest can simply be written in terms of the infallible and error-prone outcome classifiers instead of the exposure classifiers.

Greenland (2008) [13] gave a critical review of the so-called 'direct' and maximum likelihood (ML) approaches. The direct approaches rely on what is referred to in the literature as 'matrix estimation' (see references Barron (1977) [2], Marshall (1990) [28], Morrissey and Spiegelman (1999) [32]). These methods yield MCB-adjusted estimates of the cell counts or binomial success probabilities in which sample estimates are 'plugged' into formulas to replace population parameters. These authors presented the matrix form of a set of equations that define the relationship between the counts/probabilities with and without misclassification. These equations can be written in terms of either the sensitivity (SE) and specificity (SP) (known as the matrix method) or the positive predictive value (PPV) and negative predictive value (NPV) (known as the inverse matrix method). For details of these equations, see Morrissey and Spiegelman (1999) [32]. It should be noted, however, that the SP/SE matrix method is the least efficient of the two approaches (Greenland (1988) [12], Lyles (2002) [23], Greenland (2008) [13]). The plug-in estimators used to derive the estimates of the MCB-adjusted cell counts/probabilities are observed proportions obtained directly from the data. Since we have two sources of data, one can consider the validated data alone (drawn randomly for these matrix approaches), or combine it with the information from the original data (also known as 'double sampling'; see Tennebein (1970) [38]). Both of these choices have been investigated in the literature (Marshall (1990) [28], Morrissey and Spiegelman (1999) [31]), with the doubly sampled approach referred to as the 'improved' version of the matrix methods. Finally, Greenland (1988 [12], 2008 [13]) introduced a weighted estimator that considers inverse variance weighting (IVW) of the odds-ratio estimates derived from both the validated and unvalidated samples (without the use of double sampling). He also confirms that the IVW approach will be more efficient than the SP/SE matrix method. The derivation of the asymptotic variance of these 'matrix' method-based estimators is given in Greenland (1988) [12]. This method has also been applied in the context of the IVW estimator and the variance formula was reproduced in Greenland (2008) [13]. We note that there appears to be a difference between the variance formulas in Greenland (1988) [12] (formula 1 used to calculate V_r in example 4) and Greenland (2008) [13] (formula 6; v_U). This was likely due to an algebraic error when rewriting the formula in a different notation in the 2008 paper. We provide the corrected formula here.

Morrissey and Spiegelman (1999) [32] and Lyles (2003) [24] discussed the ML approaches, and defined the likelihood parametrized by SP/SE, and PPV/NPV. The latter author posited that the PPV/NPV parametrization is equivalent to the inverse matrix method (Marshall (1990) [28]). He demonstrated this by showing that the MLEs are the same as the point estimates derived in Marshall (1990) [28]. He then pointed out that the asymptotic variance formula derived for the inverse matrix method (using the approach described in Greenland (1988) [12]) is the same as for the ML approach. Next, he noted that the derivation of the MLEs under the SP/SE parametrization is "messy and requires numerical methods". However, we will demonstrate in Chapter 2 that using an appropriately chosen transformation, we are able to derive closed form expressions for the MLEs of π, SE and SP. As expected, the resulting MLEs under this likelihood are equivalent to the MLEs derived under the PPV/NPV parametrization. The validation sample contribution to the likelihood under both parameterizations are written in terms of a random sampling approach and do not consider a formulation that allows for conditional sampling based on the observed outcomes (or exposures) from the original sample. Thus, as previously mentioned, we shall consider a multi-sample framework to model the combined information in the two samples. The details of this framework are summarized in Appendix A and are adapted from work done by Hirose (2005) [17] and results presented in Lehmann and Casella (1998) [21].

Next, we review the literature regarding binary outcome misclassification and logistic regression with the use of internal validation information. Magder et al.

(1997) [27] discussed the general problem under known misclassification rates and presented an approach to derive the MLEs of the regression coefficients. They recommended an expectation-maximization (EM) approach where the misclassification rates can be estimated using information from the original and validation samples at each step. Carroll et al. (2006) [4] presented the general form of a likelihood with the presence of internal validation data. However, this general likelihood is formulated around a randomly drawn validation sample, ignoring the observed information from the original sample (see Section 15.4.2 in Carroll et al. (2006) [4]).

Lyles et al. (2011) [25] conducted an overview of likelihood based approaches to address outcome misclassification assuming known and unknown misclassification rates with the use of external and internal validation data. They applied the model described in Carroll et al. (2006) [4] with internal validation data and non-differential misclassification and presented an example considering bacterial vaginosis status in women enrolled in the HIV Epidemiology Research Study. Next, they conducted a simulation study designed to emulate this real data example. Using the regression parameter estimates observed in the example as the simulation parameter values, they demonstrated the success of the methodology. However, the likelihood presented is based on a random sampling approach that ignores the observed information from the original sample when drawing a validation sample. As such, we shall once again apply the multi-sample framework to investigate the conditional sampling approach with a logistic regression model.

Finally, as mentioned in Section 1.1, EHR data has a time-to-event structure and we have discussed misclassification of the outcome assuming that a time interval has been fixed and the response has been dichotomized in this interval. However, this ignores a great deal of temporal information, particularly the manner in which ADR risk fluctuates over time. Thus, we shall review the literature on outcome misclassification in survival data.

The literature on discrete time survival models is extensive. Richardson et al. (2000) [34] considered right-censored survival data in which the observational time points were predetermined. Individuals would be observed until the outcome of interest was recorded, in this case a positive test for an infectious disease. They developed two EM algorithms to provide product limit estimation of the survival function under the assumptions of curable and incurable diseases, respectively, under the possible presence of erroneous diagnostic testing. Meier et al. (2003) [30] extended the discrete proportional hazards model to account for measurement error assuming known observational times. Unlike Richardson et al. (2000) [34] they included covariate effects and were able to produce unbiased estimators for known misclassification rates. Finally, Magaret (2008) [26] introduced a general model for the use of validation data for the discrete proportional hazards model under misclassification and unknown error rates.

The literature in this field is more limited under a continuous time framework. McKeown et al. (2010) [29] considered estimation of the distribution function of current status data with the presence of misclassification under a continuous time framework. Note, that, we have access to more information than simply the relationship of the time to AE occurrence to some independent monitoring time. Although this is not of direct interest, we mention this as it is the only work on outcome misclassification in continuous time of which we are aware. In the context of competing risks under a C-S hazards formulation (Kalbfleisch and Prentice (2002) [19]) the problem of misclassification of C-S event type has been addressed in the literature. Rompaye et al. (2010) [36] outline a partial likelihood approach to this problem under the assumption of known misclassification rates and proportional C-S baseline hazards. They only allow for misclassification of the C-S event types and observe the time to event occurrence without error. We expand upon this model in Chapter 5 allowing for unknown error rates by incorporating validation data as well as allowing for misclassification of the event time under certain assumptions.

1.3 Thesis Objectives

This thesis is centered around an internal validation sampling approach to produce unbiased estimation for binary and time-to-event data in the presence of outcome misclassification.

- We will demonstrate that the use of the possibly error-prone observed outcome information can influence the validation sampling scheme in such a way that it is possible to produce unbiased estimators of the parameters of interest which may be more efficient and less costly than the standard approaches which ignore this additional information.
- Through the use of Monte Carlo methods, we will evaluate the influence of the validation sampling schemes on the resulting estimates of asymptotic standard error
- We introduce a Monte Carlo sample size determination approach to allow investigators to incorporate this information in the study design.
- We will apply these approaches to adjusting for MCB in the odds-ratio and logistic regression coefficients, motivated by the problem of outcome misclassification in electronic health records data.
- For continuous time right censored survival data with and without competing risks, we will address two forms of misclassification. Specifically, we consider the problem of failing to observe the event of interest and erroneously observing a censorship event followed by observing the incorrect failure event type.
- We will employ a validation sampling approach to produce unbiased risk estimates with the presence of either or both of these forms of misclassification.

Our results are theoretically justified from an asymptotic (large sample) perspective, with the statistical properties of the bias-adjusted estimators investigated through Monte Carlo simulation.

Chapter 2

On the Optimization of a Validation Sampling Approach for Unbiased Estimation in Binary Data in the Presence of Outcome Misclassification

2.1 Introduction

In this chapter, we address the problem of binary outcome misclassification in the estimation of the odds-ratio for a 2 × 2 table, extending the literature review in Section 1.2. We begin by introducing the binomial likelihood with outcome misclassification under the specificity/sensitivity (SP/SE) parametrization in Section 2.2. We then introduce the random and conditional approaches to drawing a validation sample, and present the full likelihoods under both approaches. Recall that we wish to investigate the impact of altering the underlying categorical make-up of the validation sample on the asymptotic relative efficiency (ARE). For instance, we could choose to select a validation sample with 40% category 1 and 60% category 2, or alternatively one with 20% category 1 and 80% category 2. We will demonstrate in Section 2.3 that there will be a difference in the estimated AREs of the resulting maximum likelihood estimates (MLEs), as well as the estimator based on the random sampling approach. Additionally, the random sampling approach is expected to draw validation observations at a rate that is similar to the observed incidence rates of each category in the original sample. This can be problematic for applications in which the observed incidence rates are small or a difference in validation sampling costs between categories leads to a preference for minimization of a certain category's sample size. For example, there are many rare diseases that would be of interest to investigators analyzing electronic health records (EHRs) and validating a diagnosis may require the use of expensive laboratory tests (whereas rejecting a diagnosis may be straightforward on clinical grounds). We shall investigate both the random and conditional validation sampling approaches under the SP/SE parametrization and provide the resulting closed form MLEs in Section 2.2. For the remainder of this discussion, we will refer to the random sampling approach as the 2S approach, a short form notation for two samples; the original EHR data and a single validation sample drawn randomly this dataset. The conditional sampling approach draws two validation samples, the first from the subset of the original data with observed adverse event (AE) status 1 (AE of interest), and the second from the reference subset. Thus, we shall refer to this as the 3S approach to signify a three sample structure.

In Section 2.3, we introduce the multi-sample framework underlying the data and justify its use to define misclassified binary data. We will present the asymptotic properties of the MLEs derived under this framework of multiple samples and provide a consistent estimate of the asymptotic covariance matrix for the model parameters. Section 2.3 also includes a simulation study designed to numerically investigate the asymptotic properties of these methods and compare the maximum likelihood (ML) approaches. Section 2.4 discusses the inverse variance weighted (IVW) matrix estimator and presents the algebraic correction to the variance formula in Greenland (2008) [13] (mentioned in Section 1.2). Simulations conducted to compare the IVW estimator with the 2S and 3S approaches are also presented.

2.2 Binomial Likelihood with Outcome Misclassification

In this section, we introduce the binomial likelihood with outcome misclassification under the SP/SE parametrization using notation that is based on the EHR data example discussed in Chapter 1. Consider two binary classifiers, one for the outcome of interest, A, and the other, D to signify drug exposure group. Specifically, A takes on a value of one under the presence of some outcome of interest (AE), and zero otherwise. Similarly D assumes the value of one under the presence of some exposure (drug) of interest, and zero otherwise. We are interested in the probabilities of classifying $A = 1$ in both exposure groups, $d = 0, 1$. Thus, we define this probability as $\pi_{1|d} = P(A = 1|D = d)$, allowing us to describe the source population in the form of a 2 x 2 contingency table given in Table 2.2.1 with cell counts denoted by $N_{da} = \sharp\{D = d, A = a\}$.

Table 2.2.1: Cell probabilities in the 2x2 cross-classification of the source population.

	$A = 1$	$A = 0$
$D = 1$	$\pi_{1\|1}$	$\pi_{0\|1}$
$D = 0$	$\pi_{1\|0}$	$\pi_{0\|0}$

Assume that D is classified without error, but instead of observing A, we obtain a possibly error-prone classifier, $\tilde{A}$, where $\tilde{A} = 1$ if we observe the possibly misclassified presence of the outcome of interest, and zero otherwise. Then, we can write the misclassification probabilities as,

$$\theta_{da} = P[\tilde{A} = 1|D = d, A = a], \tag{2.1}$$

where $d = 0, 1$ and $a = 0, 1$.

Note that using this notation, SE and SP are represented by θ_{d1} and $1 - \theta_{d0}$ respectively, $d = 0, 1$. Thus, we are able to write the cell probabilities underlying the observed data as

$$\begin{aligned} P[\tilde{A} = 1 | D = d] &= \theta_{d1}\pi_{1|d} + \theta_{d0}\pi_{0|d} \\ P[\tilde{A} = 0 | D = d] &= 1 - \theta_{d1}\pi_{1|d} - \theta_{d0}\pi_{0|d}, \end{aligned} \tag{2.2}$$

for $d = 0, 1$. The 2 x 2 contingency table for the probabilities of the observed data is given in Table 2.2.2 with cell counts denoted by $n_{d\tilde{a}} = \sharp\{D = d, \tilde{A} = \tilde{a}\}$, $d, \tilde{a} = 0, 1$.

Table 2.2.2: Cell probabilities in the 2x2 cross-classification of the observed data.

	$\tilde{A} = 1$	$\tilde{A} = 0$
$D = 1$	$\theta_{11}\pi_{1\|1} + \theta_{10}\pi_{0\|1}$	$1 - \theta_{11}\pi_{1\|1} - \theta_{10}\pi_{0\|1}$
$D = 0$	$\theta_{01}\pi_{1\|0} + \theta_{00}\pi_{0\|0}$	$1 - \theta_{01}\pi_{1\|0} - \theta_{00}\pi_{0\|0}$

This produces the following likelihood of the observed EHR data,

$$L_O = \prod_{d=0,1} (\theta_{d1}\pi_{1|d} + \theta_{d0}\pi_{0|d})^{n_{d1}} (1 - \theta_{d1}\pi_{1|d} - \theta_{d0}\pi_{0|d})^{n_{d0}}, \tag{2.3}$$

where the number of individuals in each exposure group is $n_{d.} = n_{d1} + n_{d0}$.

Clearly, in the absence of any knowledge on the θ-parameters, the π-parameters cannot be estimated. Because we require some form of additional information, we will take an internal validation sample as previously discussed. Recall that this secondary sample is drawn from the original sample, and, unlike the original sample, is not subject to misclassification. The additional information in the validation sample allows us to estimate the θ-parameters and finally to derive misclassification bias (MCB) adjusted estimates of the π-parameters.

As mentioned in Chapter 1, selection of the validation sample can be done in two ways. Under the 2S approach, the experimenter will randomly draw a validation sample of size m from the original data and ignore the observed $\tilde{A}$ category. Alternatively, under the 3S approach, the experimenter will select a sample from each $\tilde{A}$ group, $\tilde{a} = 0, 1$. These methods of sampling possess different underlying probabilistic structures that must be considered when developing the full likelihood. Denote the validation sample size in the $d, \tilde{a}^{th}$ group by $m_{d\tilde{a}}$ and let $y_{d\tilde{a}}$ be the number of those individuals that have an updated AE status of $A = 1$ for $d, \tilde{a} = 0, 1$. Hence, the 3S approach draws subsamples of size $m_{.\tilde{a}} = m_{1\tilde{a}} + m_{0\tilde{a}}$ for each observed outcome category for a total validation sample size of $m = m_{.1} + m_{.0}$. We display the observed data from both the original and validation samples in Table 2.2.3. Note that this table is formatted to be similar to Table 1 in Greenland (2008) [13].

Table 2.2.3: Summary table of cell counts for the original and validation data.

	D=1		D=0	
	$\tilde{A} = 1$	$\tilde{A} = 0$	$\tilde{A} = 1$	$\tilde{A} = 0$
$A = 1$	y_{11}	y_{10}	y_{01}	y_{00}
$A = 0$	$m_{11} - y_{11}$	$m_{10} - y_{10}$	$m_{01} - y_{01}$	$m_{00} - y_{00}$
Unvalidated	$n_{11} - m_{11}$	$n_{10} - m_{10}$	$n_{01} - m_{01}$	$n_{00} - m_{00}$
Total	n_{11}	n_{10}	n_{01}	n_{00}

Construction of the 3S approach begins by defining the validation sample contributions to the likelihood as,

$$P[A = a | D = d, \tilde{A} = \tilde{a}] = \frac{P[\tilde{A} = \tilde{a} | D = d, A = a] P[A = a | D = d]}{P[\tilde{A} = \tilde{a} | D = d]}, \qquad (2.4)$$

for $d, a, \tilde{a} = 0, 1$.

This yields the following validation sample likelihoods for each $\tilde{a}$,

$$\begin{aligned} L_{V_1} &= \prod_{d=0,1} \left(\frac{\theta_{d1}\pi_{1|d}}{\theta_{d1}\pi_{1|d} + \theta_{d0}\pi_{0|d}} \right)^{y_{d1}} \left(\frac{\theta_{d0}\pi_{0|d}}{\theta_{d1}\pi_{1|d} + \theta_{0d}\pi_{0|d}} \right)^{m_{d1}-y_{d1}} \\ L_{V_0} &= \prod_{d=0,1} \left(\frac{(1-\theta_{d1})\pi_{1|d}}{1-\theta_{d1}\pi_{1|d} - \theta_{d0}\pi_{0|d}} \right)^{y_{d0}} \left(\frac{(1-\theta_{d0})\pi_{0|d}}{1-\theta_{d1}\pi_{1|d} - \theta_{d0}\pi_{0|d}} \right)^{m_{d0}-y_{d0}} . \end{aligned} \quad (2.5)$$

The full likelihood of the data with the presence of binary misclassification in the outcome variable is obtained by combining L_O, L_{V_1} and L_{V_0} as follows,

$$\begin{aligned} L_O \times L_{V_1} \times L_{V_0} &= \prod_{d=0,1} (\theta_{d1}\pi_{1|d} + \theta_{d0}\pi_{0|d})^{n_{d1}} (1-\theta_{d1}\pi_{1|d} - \theta_{d0}\pi_{0|d})^{n_{d0}} \qquad (2.6) \\ &\quad \times \left(\frac{\theta_{d1}\pi_{1|d}}{\theta_{d1}\pi_{1|d} + \theta_{d0}\pi_{0|d}} \right)^{y_{d1}} \left(\frac{\theta_{d0}\pi_{0|d}}{\theta_{d1}\pi_{1|d} + \theta_{0d}\pi_{0|d}} \right)^{m_{d1}-y_{d1}} \\ &\quad \times \left(\frac{(1-\theta_{d1})\pi_{1|d}}{1-\theta_{d1}\pi_{1|d} - \theta_{d0}\pi_{0|d}} \right)^{y_{d0}} \left(\frac{(1-\theta_{d0})\pi_{0|d}}{1-\theta_{d1}\pi_{1|d} - \theta_{d0}\pi_{0|d}} \right)^{m_{d0}-y_{d0}} \\ &= \prod_{d=0,1} (\theta_{d1}\pi_{1|d} + \theta_{d0}\pi_{0|d})^{n_{d1}-m_{d1}} (1-\theta_{d1}\pi_{1|d} - \theta_{d0}\pi_{0|d})^{n_{d0}-m_{d0}} \\ &\quad \times (\theta_{d1}\pi_{1|d})^{y_{d1}} (\theta_{d0}\pi_{0|d})^{m_{d1}-y_{d1}} [(1-\theta_{d1})\pi_{1|d}]^{y_{d0}} [(1-\theta_{d0})\pi_{0|d}]^{m_{d0}-y_{d0}}. \end{aligned}$$

On the other hand, the 2S approach will have a validation sample with the following underlying probabilistic structure,

$$P[A = a, \tilde{A} = \tilde{a} | D = d] = P[\tilde{A} = \tilde{a} | D = d, A = a] P[A = a | D = d], \qquad (2.7)$$

for $d, a, \tilde{a} = 0, 1$. This leads to the 2S validation sample likelihood,

$$L_V = \prod_{d=0,1} (\theta_{d1}\pi_{1|d})^{y_{d1}} (\theta_{d0}\pi_{0|d})^{m_{d1}-y_{d1}} [(1-\theta_{d1})\pi_{1|d}]^{y_{d0}} [(1-\theta_{d0})\pi_{0|d}]^{m_{d0}-y_{d0}}.$$

Since these joint probabilities account for the dependencies between the two observed classifications, the original outcome status is only represented in the likelihood once. Hence, the 2S full likelihood of the data is,

$$\begin{aligned} L_O \times L_V &= \prod_{d=0,1} (\theta_{d1}\pi_{1|d} + \theta_{d0}\pi_{0|d})^{n_{d1}-m_{d1}}(1 - \theta_{d1}\pi_{1|d} - \theta_{d0}\pi_{0|d})^{n_{d0}-d0} \\ &\times \quad (\theta_{d1}\pi_{1|d})^{y_{d1}}(\theta_{d0}\pi_{0|d})^{m_{d1}-y_{d1}}[(1-\theta_{d1})\pi_{1|d}]^{y_{d0}}[(1-\theta_{d0})\pi_{0|d}]^{m_{d0}-y_{d0}} \end{aligned} \tag{2.8}$$

Note, that these likelihoods simplify to the same expression, however, the differences in sampling lead to different probabilistic structures underlying the data. Under the 3S approach, each validation sample portion of the likelihood is based on a fixed sub sample size $m_{.1}$ and $m_{.0}$ where $m = m_{.1} + m_{.0}$. In other words, for each $\tilde{a} = 0, 1$, we have a quadrinomial distribution with the probabilities listed in equation (2.4) for $d, a = 0, 1$. Using the 2S approach, we are only able to fix the overall validation sample size, m, which leads to a multinomial distribution with the probabilities listed in equation (2.7) for $d, \tilde{a}, a = 0, 1$.

We can use the likelihood to derive closed-form expressions for the MLEs of both the π and θ-parameters. Consider the transformations, $\phi_{1|d} = \pi_{1|d}\theta_{d1}$ and $\phi_{0|d} = \pi_{0|d}\theta_{d0}$. This allows us to write the likelihood as,

$$\begin{aligned} L &= \prod_{d=0,1} (\phi_{1|e} + \phi_{0|d})^{n_{d1}-m_{d1}}(1 - \phi_{1|d} - \phi_{0|d})^{n_{d0}-m_{d0}} \\ &\quad \times \phi_{1|d}^{y_{d1}}\phi_{0|d}^{m_{d1}-y_{d1}}(\pi_{1|d} - \phi_{1|d})^{y_{d0}}(\pi_{0|d} - \phi_{0|d})^{m_{d0}-y_{d0}}. \end{aligned} \tag{2.9}$$

Maximizing $\log(L)$ with respect to the transformed parameters gives the corresponding ϕ-parameter MLEs, $\hat{\phi}_{1|d} = \frac{y_{d1}n_{d1}}{m_{d1}(n_{d1}+n_{d0})}$ and $\hat{\phi}_{0|d} = \frac{n_{d1}(m_{d1}-y_{d1})}{m_{d1}(n_{d1}+n_{d0})}$ and the MLEs of interest are,

$$\begin{aligned} \hat{\pi}_{1|d} &= \frac{m_{d1}n_{d0}y_{d0} + m_{d0}n_{d1}y_{d1}}{m_{d0}m_{d1}(n_{d1}+n_{d0})} \\ \hat{\theta}_{d1} &= \frac{m_{d0}y_{d1}n_{d1}}{m_{d0}y_{d1}n_{d1} + m_{d1}y_{d0}n_{d0}} \\ \hat{\theta}_{d0} &= \frac{m_{d0}n_{d1}(m_{d1} - y_{d1})}{m_{d0}y_{d1}n_{d1} + m_{d1}y_{d0}n_{d0}}. \end{aligned} \tag{2.10}$$

Since both the 2S and 3S approach share the common likelihood, equation (2.9), these MLEs apply to both sampling methods. Also note that these MLEs correspond to those presented in Lyles (2002) [23].

Next, we will consider some of the issues discussed in Section 2.1 regarding the ARE of these estimators. In particular, we will compare the 2S and 3S approaches while altering the size of $\frac{m_{.1}}{m}$. Hence, in the next section, we first introduce the multi-sample framework and present the asymptotic properties of the above estimators, followed by a small simulation study to investigate the estimated AREs.

2.3 A Multi-Sample Framework for MCB-Adjusted Estimation

A formal definition for a multi-sample framework is described in Appendix A. Using this description, we shall analogously define the data under both the 2S and 3S approaches. Consider drawing n observations from a source population where m of them are measured twice, first by the fallible classifier, and then by the infallible classifier. The remaining unvalidated observations, $n^u = n - m$, are only measured by the fallible classifier. For the 2S approach, we can write the data as $(A_1, \tilde{A}_1, D_1), ..., (A_m, \tilde{A}_m, D_m), (\tilde{A}_{m+1}, D_{m+1}), ..., (\tilde{A}_n, D_n)$ where the sample proportions are $(\frac{m}{n}, \frac{n^u}{n})$, which clearly sum to 1. Since the original EHR data is summarized in a 2×2 contingency table (see Table 2.2.2) conditional on exposure group, it can be described by two binomial distributions with parameters, $P(\tilde{A} = 1|D = d)$ and $n_{d.}$ for $d = 0, 1$. The 2S approach sorts the validation data into eight categories that can therefore be described by a multinomial distribution with cell probabilities, $P(A = a, \tilde{A} = \tilde{a}|D = d)$, $a, \tilde{a} = 0, 1$ for each $d = 0, 1$.

For the 3S approach, we have a similar set up, except that the m observations are doubly measured relative to the infallible classifier's observed realizations. Thus, the data can be written as, $(A_1, \tilde{A}_1, D_1), ...,$ $(A_{m_{.1}}\tilde{A}_{m_{.1}}, D_{m_{.1}})$, $(A_{m_{.1}+1}, \tilde{A}_{m_{.1}+1},$

$D_{m_{.1}+1}), ..., (A_{m_{.1}+m_{.0}}, \tilde{A}_{m_{.1}+m_{.0}}, D_{m_{.1}+m_{.0}}), (\tilde{A}_{m+1}, D_{m+1}), ..., (\tilde{A}_n, D_n)$ where the sample proportions are $(\frac{m_{.1}}{n}, \frac{m_{.0}}{n}, \frac{n^u}{n})$ which sum to 1 and $m = m_{.1} + m_{.0}$. The original EHR data, under this approach, is described in the same manner as the 2S approach. However, the validation sample now conditions on both the exposure group as well as the observed outcome status, $\tilde{a}$. Thus, we can describe this data as being quadrinomially distributed with success probabilities $P(A = 1|\tilde{A} = \tilde{a}, D = d)$ for each $\tilde{a}, d = 0, 1$.

Thus, the definition of multi-sample data as described in Definition A.2.1 in Appendix A is satisfied and we are able to apply the results of Theorem A.2.3. Given that the data structure associated with each individual sample is described as arising from either a binomial or multinomial distribution the required regularity conditions are satisfied. Thus, we have,

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0) \xrightarrow{D} N\left(\mathbf{0}, \left[\sum_{s=1}^{s} \omega_s \mathcal{I}_s(\gamma_0)\right]^{-1}\right), \tag{2.11}$$

where γ denotes the vector of parameters, $\gamma = (\pi_{1|1}, \pi_{1|0}, \theta_{11}, \theta_{01}, \theta_{10}, \theta_{00})^T$, $\hat{\gamma}$ denotes the MLE which tends to the hypothesized value, γ_0, in probability, ω_s denotes the s^{th} probability weight in Definition A.2.1 and $\mathcal{I}_s(\gamma_0)$ is the negative expectation of the Hessian matrix for the s^{th} sample at the hypothesized value, γ_0, $s = 1, ..., S$.

Estimation of this asymptotic covariance matrix is accomplished using Theorem A.2.4, which allows us to write,

$$\frac{1}{n}\sum_{s=1}^{S} n_s \bar{I}_s(\hat{\gamma}) \xrightarrow{p} \sum_{s=1}^{S} \omega_s \mathcal{I}_s(\gamma_0), \tag{2.12}$$

where $\bar{I}_s(\hat{\gamma}) = \frac{1}{n_s}\sum_{i=1}^{n_s} I_{si}(\hat{\gamma}) = \frac{1}{n_s}\sum_{i=1}^{n_s}\left[-\frac{\partial l_s(\gamma|x_{si})}{\partial\gamma\partial\gamma^T}\right]_{\gamma=\hat{\gamma}}$ and $n = \sum_{s=1}^{S} n_s$. Note that $I(\hat{\gamma}) = \frac{1}{n}\sum_s^S \sum_i^{n_s} I_{si}(\hat{\gamma})$, which implies that the inverse of the observed information matrix of the overall likelihood is a consistent estimator of the asymptotic covariance matrix.

These results may be used to conduct a preliminary simulation study to compare the 2S and 3S approaches. As stated in Section 2.1, we wish to observe the effect of altering the categorical make-up of the validation sample on the ARE between the resulting MLEs. To ensure that the results can be compared with those in the literature, we will focus on the odds-ratio,

$$OR = \frac{\pi_{1|1}(1-\pi_{1|0})}{\pi_{1|0}(1-\pi_{1|1})} \tag{2.13}$$

and will apply the delta method in the following manner. Since (2.13) can be written as,

$$\log OR = g(\gamma) = \log \frac{\pi_{1|1}}{1-\pi_{1|1}} - \log \frac{\pi_{1|0}}{1-\pi_{1|0}},$$

we can write,

$$\begin{aligned} &\hat{\sigma}^2(\log \widehat{OR}) && (2.14)\\ &= \left(\left.\frac{\partial g}{\partial \pi_{1|1}}\right|_{\hat{\gamma}}, \left.\frac{\partial g}{\partial \pi_{1|0}}\right|_{\hat{\gamma}}, 0,0,0,0 \right)^T [I(\hat{\gamma})]^{-1} \left(\left.\frac{\partial g}{\partial \pi_{1|1}}\right|_{\hat{\gamma}}, \left.\frac{\partial g}{\partial \pi_{1|0}}\right|_{\hat{\gamma}}, 0,0,0,0 \right), \end{aligned}$$

where $\frac{\partial g}{\partial \pi_{1|1}} = \frac{1}{\pi_{1|1}(1-\pi_{1|1})}$, $\frac{\partial g}{\partial \pi_{1|0}} = -\frac{1}{\pi_{1|0}(1-\pi_{1|0})}$, $I(\hat{\gamma})$ is the overall observed information matrix and σ^2 is used to denote the variance.

Equation (2.14) provides the estimated asymptotic variance of the estimator of the log odds-ratio for both the 2S and 3S approaches at the point of their respective MLEs. Thus, under the 2S approach, we can write,

$$I_2(\hat{\gamma}) = n^u \bar{I}_O(\hat{\gamma}) + m\bar{I}_V(\hat{\gamma}), \tag{2.15}$$

where the O subscript denotes the original sample and the V subscript denotes the validation sample.

Next, we wish to investigate the effect of altering the categorical make-up of the validation sample under the 3S approach. To do so we will introduce the notation,

$$P_1 = \frac{m_{.1}}{m}, \tag{2.16}$$

to denote the chosen proportion of category 1 observations in the validation sample. This implies that, for the 3S approach, the observed information matrix, $I_3(\hat{\gamma})$ can be written as,

$$I_3(\hat{\gamma}) = n\bar{I}_O(\hat{\gamma}) + m_{.1}\bar{I}_{V_1}(\hat{\gamma}) + m_{.0}\bar{I}_{V_0}(\hat{\gamma}) = n\bar{I}_O(\hat{\gamma}) + m[P_1\bar{I}_{V_1}(\hat{\gamma}) + (1-P_1)\bar{I}_{V_0}(\hat{\gamma})], \tag{2.17}$$

where $V_{\tilde{a}}$ denotes the validation sample associated with observations of type $\tilde{A} = \tilde{a}$, $\tilde{a} = 0, 1$. This form provides the ability to simulate across a range of values of $P_1 \in [0, 1]$. In Appendix C we present the R code that can be used to compute the $\bar{I}$'s in both $I_2(\hat{\gamma})$ and $I_3(\hat{\gamma})$.

2.3.1 Comparative Simulation Studies: 2S versus 3S Approaches

In the simulation study, we compare the 2S and 3S approaches by selecting the target simulation parameters, $\gamma = (\pi_{1|1}, \pi_{1|0}, \theta_{11}, \theta_{01}, \theta_{10}, \theta_{00})^T$, and applying the ML methods to produce MCB-adjusted estimates of the oods-ratio in a 2×2 table. First, we must consider how to generate binary data with misclassification of the outcome.

We begin with the following chosen values: N_d, θ_{da} and $\pi_{a|d}$ for $d, a = 0, 1$. To generate N_{da} (the target drug/AE counts), we will take a random binomial draw from N_d with probability $\pi_{1|d}$. To generate the $n_{d\tilde{a}}$'s, each N_{da} with possible misclassification can be split into two groups: the correctly classified observations, which will be denoted CC_{da}, and the misclassified observations, which will be denoted MC_{da}. The final step is to determine how to choose the number of observations that were correctly or incorrectly classified from each N_{da} group.

Note that $n_{d1} = CC_{d1} + MC_{d0}$ and $n_{d0} = CC_{d0} + MC_{d1}$ for $d = 0, 1$. We define the random quantities CC_{da} to be binomially distributed with the number of trials to be N_{da}. The associated probability parameters are chosen to ensure that the $n_{d\tilde{a}}$'s are generated using the structure of the data defined in Table 2.2.2. This can be accomplished by defining the success probability for CC_{d1} as θ_{d1} and $1 - \theta_{d0}$ for CC_{d0} giving us the associated probabilities needed to generate our observations as described in Table 2.2.2. R-code to generate this data is included in Appendix C.

Once the data are generated, we select a validation sample by either selecting a random sample of size m (2S approach), or by drawing two validation samples of size $m_{.1}$ and $m_{.0}$ from the subsets of the original sample associated with $\tilde{A} = 1$ and 0 respectively. Once these samples have been obtained, we can observe the necessary information is available to compute the odds-ratio estimates under both approaches, $\widehat{OR} = \frac{\hat{\pi}_{1|1}(1-\hat{\pi}_{1|0})}{\hat{\pi}_{1|0}(1-\hat{\pi}_{1|1})}$, where the π-parameter estimates are as described in Section 2.2.

Since the validation sample will be drawn from the original data in practice, we will have access to the observed incidence rate, $\frac{n_{.1}}{n}$ where $n_{.\tilde{a}}$ denotes the number of observations in the original EHR data associated with category $\tilde{a} = 0, 1$ ($n_{.\tilde{a}} = n_{1\tilde{a}} + n_{0\tilde{a}}$). As such, in generating data for this simulation we shall pay careful attention to these rates, as they will influence the resulting categorical make-up of the validation sample in the 2S approach. Thus, we choose simulation parameters based on chosen expected counts of $\tilde{A} = 1$. The expected count can be thought of as the sum of the products of the exposure group sizes and their respective conditional probabilities, $n_{1.}P(\tilde{A} = 1|D = 1) + n_{0.}P(\tilde{A} = 1|D = 0)$ where $n_{d.}$ denotes the exposure group (drug group if referring to the original example) sample size in the original data. Note that for this preliminary simulation study, we fix the exposure group sizes to be equal throughout this section. We do so since we are primarily interested in the ARE relative to the $\tilde{A}$ groups for this study. The major concern when altering the exposure groups will be the effect on estimation of inadequate subgroup sample sizes; however, this will affect both the 2S and 3S approaches in the same manner. Nevertheless, we shall investigate this point in more detail in the simulations summarized in Section 2.4.1. Thus, we can write the expected number of category 1 observations in the original sample as $\frac{n}{2}(\pi_{1|1}\theta_{11} + (1 - \pi_{1|1})\theta_{10} + \pi_{1|0}\theta_{01} + (1 - \pi_{1|0})\theta_{00})$. Note that in Table 2.3.1 we present these values in the column marked $E(n_{.1})$.

Table 2.3.1: Chosen target parameters for the simulation studies comparing the 2S and 3S approaches.

ID	$E(n_{.1})$	$\pi_{1\|1}$	$\pi_{1\|0}$	θ_{11}	θ_{01}	θ_{10}	θ_{00}
1	$0.097n$	0.05	0.10	0.85	0.95	0.05	0.01
2	$0.195n$	0.1	0.20	0.95	0.85	0.05	0.10
3	$0.300n$	0.3	0.20	0.90	0.90	0.10	0.10

In Table 2.3.1, we have selected three sets of simulation parameters designed to slowly increase the incidence rates. However, note that these are a subset of a larger group of simulation studies, all producing similar results and as such their results were omitted (we display the additional results in Appendix D). Using these parameter sets, we generate estimates of the odds-ratio, the sample standard deviation of these estimates, delta method estimates of standard error, and estimates of coverage proportion based on 95% confidence intervals for both 2S and 3S approaches. We also report the analogous estimates based only on the observed data, ignoring the possibility of misclassification. Note that the odds-ratio estimates with perfect classification are calculated in the standard way, $\widetilde{OR} = \frac{n_{11}n_{00}}{n_{10}n_{01}}$, with standard error estimates, $\hat{\sigma}(\log \widetilde{OR}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{01}} + \frac{1}{n_{10}} + \frac{1}{n_{00}}}$ (Agresti (2002) [1]). The tilde is used to identify these estimates. In this simulation, we considered different combinations of sample sizes, $n = (10,000, 100,000)$ and $m = (0.01n, 0.05n, 0.1n)$ as these are realistic values for EHRs. Finally, each simulation was conducted over 1000 iterations.

As the results are similar across all parameter sets as well as the values of n and m, we only report a few sets of results; specifically those associated with $n = 10,000$, $m = 1,000$. For each parameter set, we first display the standard error estimates associated with the π-parameters across eight values of P_1. Tables 2.3.2a, 2.3.3a and 2.3.4a display the estimates for the 2S approach on the left hand column followed by the eight estimates of standard error for $\hat{\pi}_{1|1}$ and $\hat{\pi}_{1|0}$. The last row is the maximum between these two estimates. The odds-ratio tables report all the estimates mentioned above, but across three different methods. The first row are the results that ignore misclassification, followed by the 2S approach. Next, we present the 3S approach with chosen values of P_1 corresponding to the value of P_1 associated with the minimum value of the $\max(\hat{\sigma}(\hat{\pi}_{1|1}), \hat{\sigma}(\hat{\pi}_{1|0}))$ row in the previous table followed by the observed incidence rate in the original sample.

Simulation study results for the 2S versus 3S approach for parameter set 1.

(a) Monte Carlo estimates of standard error.

		3S: P_1							
	2S	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
$\hat{\sigma}(\hat{\pi}_{1\|1})$	0.00795	0.00787	0.00655	0.00616	0.00622	0.0065	0.00722	0.0085	0.01129
$\hat{\sigma}(\hat{\pi}_{1\|0})$	0.00643	0.00649	0.00597	0.00597	0.00617	0.00649	0.00718	0.0084	0.01107
max	0.00795	0.00787	0.00655	0.00616	0.00622	0.0065	0.00722	0.0085	0.01129

(b) Odds-ratio summary table (OR = 0.4737).

Approach	Average Point Estimate	Average Standard Error	Sample Standard Deviation	CP
$\widetilde{OR}$	0.853	0.0677	0.0676	0.000
$\widehat{OR}_2$	0.476	0.1841	0.1864	0.953
$\widehat{OR}_3^{min}$	0.474	0.1450	0.1386	0.963
$\widehat{OR}_3^{in}$	0.473	0.1844	0.1837	0.957

Table 2.3.2: Simulation study results for the 2S versus 3S approach for parameter set 2.

(a) Monte Carlo estimates of standard error.

		3S: P_1							
	2S	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
$\hat{\sigma}(\hat{\pi}_{1\|1})$	0.00904	0.01178	0.00906	0.00793	0.0075	0.0074	0.00768	0.00849	0.01046
$\hat{\sigma}(\hat{\pi}_{1\|0})$	0.01342	0.01659	0.01334	0.0124	0.01209	0.01233	0.01299	0.01408	0.01682
max	0.01342	0.01659	0.01334	0.0124	0.01209	0.01233	0.01299	0.01408	0.01682

(b) Odds-ratio summary table (OR = 0.4444).

Approach	Average Point Estimate	Average Standard Error	Sample Standard Deviation	CP
$\widetilde{OR}$	0.49139	0.052217	0.050818	0.522
$\widehat{OR}_2$	0.45012	0.13127	0.12427	0.969
$\widehat{OR}_3^{min}$	0.45476	0.11148	0.1116	0.936
$\widehat{OR}_3^{in}$	0.45132	0.1305	0.12667	0.968

Table 2.3.3: Simulation study results for the 2S versus 3S approach for parameter set 3.

(a) Monte Carlo estimates of standard error.

		3S: P_1							
	2S	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
$\hat{\sigma}(\hat{\pi}_{1\|1})$	0.01376	0.01986	0.01543	0.01393	0.01309	0.0128	0.01341	0.01382	0.01635
$\hat{\sigma}(\hat{\pi}_{1\|0})$	0.01275	0.01923	0.01487	0.01285	0.01215	0.01166	0.01174	0.01199	0.01397
max	0.01376	0.01986	0.01543	0.01393	0.01309	0.0128	0.01341	0.01382	0.01635

(b) Odds-ratio summary table (OR = 1.7143).

Approach	Average Point Estimate	Average Standard Error	Sample Standard Deviation	CP
$\widetilde{OR}$	1.4671	0.043972	0.044676	0.062
$\widehat{OR}_2$	1.7436	0.10439	0.10839	0.938
$\widehat{OR}_3^{min}$	1.7357	0.097327	0.080904	0.995
$\widehat{OR}_3^{in}$	1.7515	0.10514	0.11583	0.97

The results obtained demonstrate several important points. First, the delta method estimates of the asymptotic standard error, denoted as "Average Standard Error" in Tables 2.3.2b, 2.3.3b and 2.3.4b, are a good approximation of the sample standard deviation of the $\widehat{OR}$ estimates, denoted as "Sample Standard Deviation"; however, note that for small values of m, the approximation worsens. Next, as expected, the point estimates displayed in the column marked "Average Point Estimates", under both 2S and 3S approaches do quite well. Finally, the column marked "CP" are estimates of coverage proportion for the odds-ratio estimates ignoring MCB in the $\widetilde{OR}$ row and for those adjusting for MCB in the remaining rows. Clearly, the estimates associated with $\widetilde{OR}$ are performing poorly while all three of the adjusted methods do quite well as the nominal coverage rate is set at 95%.

Next, considering Tables 2.3.2a, 2.3.3a and 2.3.4a, the standard errors vary with respect to P_1 in the 3S approach. It seems as though, for values of P_1 close to 0 or 1, the 2S approach has lower variance, while for validation samples with mid-range values of P_1 the 3S approach has smaller variance. The transition point with respect to P_1, in which the 3S approach starts producing smaller estimates of standard error, appears to be related to the observed incidence rate of category 1 observations in the original data, $\frac{n_{.1}}{n}$. Parameter set 1 has an incidence rate of approximately 0.097 and the transition appears to occur close to $P_1 = 0.1$. The estimate of standard error based on the 2S approach is 0.00795 while the 3S approach with $P_1 = 0.1$ has an estimated standard error of 0.00787 and 0.00655 at $P_1 = 0.2$. These results are similar for parameter set 2 (with an incidence rate of 0.195). The 2S estimate is 0.01342 and the transition appears to occur close to $P_1 = 0.2$ with a 3S standard error estimate of 0.01334. Again, for parameter set 3 (an incidence rate of 0.3) it occurs close to $P_1 = 0.3$ with the 2S estimate at 0.01376 and the 3S estimate at 0.01393. These results make sense as under the 2S approach $\hat{m}_{.1}$ is a random variable that can be thought of as binomially distributed with parameters, m and $\frac{n_{.1}}{n}$. Thus, the likely range of observed realizations of $\hat{m}_{.1}$ is $m\frac{n_{.1}}{n} - z_{\alpha/2}\sqrt{\frac{n_{.1}n_{.0}}{n^2 m}}, ..., m\frac{n_{.1}}{n} + z_{\alpha/2}\sqrt{\frac{n_{.1}n_{.0}}{n^2 m}}$, where $n_{.\tilde{a}}$ is the number of observations with classification $\tilde{a}$ in the original sample, $\tilde{a} = 0, 1$. In other words, we would expect to see similar results as the validation samples are likely similar in composition.

Finally, we have numerically demonstrated that the 2S and 3S approaches will produce MCB-adjusted estimates of the odds-ratio, and that the more efficient method is the 3S approach under certain values of P_1. The ideal values of P_1 were 0.2 for the 1st parameter set, 0.4 for the second and 0.5 for the third. However, recall that the estimates reported here were only calculated at the eight values of P_1 and the actual ideal value may be slightly above or below. To further investigate this, we divide the interval into eighty smaller sections, simulate for each value of P_1 and report the standard error. Plotting each of these along with the value of the standard error under the 2S approach (the horizontal line) produces the following displays for parameter sets 1 and 3.

Figure 2.3.1: Plot of Monte Carlo estimates of $\max(\hat{\sigma}(\hat{\pi}_{1|1}), \hat{\sigma}(\hat{\pi}_{1|0}))$ versus P_1 for parameter set 1.

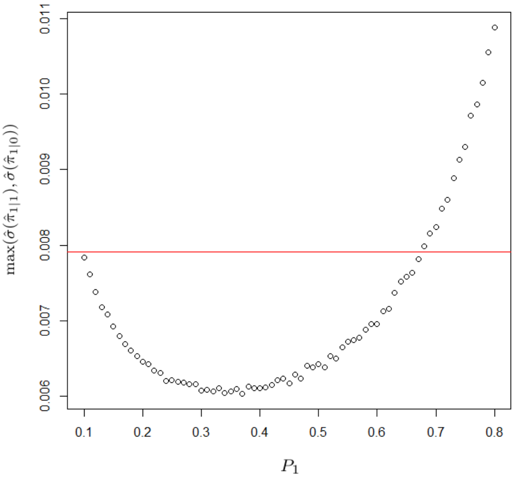
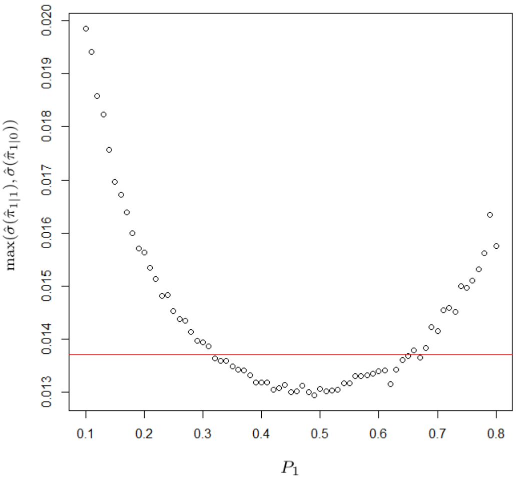


Figure 2.3.2: Plot of Monte Carlo estimates of $\max(\hat{\sigma}(\hat{\pi}_{1|1}), \hat{\sigma}(\hat{\pi}_{1|0}))$ versus P_1 for parameter set 3.



Both examples produce parabolic shaped curves for the 3S approach across the range of values of P_1. The intersection of these curves with the 2S standard error gives us an interval in which the 3S approach outperforms the 2S approach. Based on these results, it appears that the lower bound of that interval is approximately the incidence rate of category 1 in the original sample (0.097 for parameter set 1 and 0.3 for parameter set 3). Note that when the incidence rate exceeds 0.5, this relationship flips and the upper bound of the interval will approximate the underlying incidence rate. However, in both cases, the alternate bound does not appear to follow any relationship with $\frac{n_{.1}}{n}$.

Through this simulation, we have demonstrated that both the 2S and 3S approaches produce unbiased estimates of OR while $\widetilde{OR}$ is not unbiased. Further, the 3S approach with a validation sample of a particular categorical make-up will produce more efficient estimation for these parameter sets. It appears that the ideal value of P_1 is in a range of values for which one of the bounds is a function of the observed incidence rate of category 1 observations. In the next section, we investigate these intervals in a larger scale simulation study. While the proposed 3S approach under the ideal P_1 conditions may prove to be a better estimator than the 2S approach much of the time, this is not necessarily the case in general. We will also introduce Greenland's IVW estimator (Greenland (1988) [12], Greenland (2008) [13]) and investigate the ARE between both the 2S and 3S approaches and the IVW approach.

2.4 The Inverse Variance Weighted Estimator of the Odds-Ratio

In this section, we discuss the work of Greenland (1988) [12] and Greenland (2008) [13] with respect to our goal of conducting a simulation study to numerically compare the estimated ARE between the IVW estimator and the 2S and 3S approaches. To accomplish this goal, we first consider the variance formula (equation 6) of Greenland (2008) [13]. As stated in Section 2.1, we believe there to be a typographical error, particularly when comparing the formula to the work in Greenland (1988) [12]. Following this discussion, we conduct the comparative simulation study with the updated variance formula. We develop the section around outcome misclassification using our notation as described in Table 2.2.3; any additional notation borrowed from Greenland (1988, 2008) is used only in this section.

The IVW estimator (Greenland (2008) [13]) calculates two log odds-ratio estimates, one from the validation data,

$$\log \widehat{OR}_V = \log \left[\frac{(y_{11} + y_{10})(m_{01} - y_{01} + m_{00} - y_{00})}{(y_{01} + y_{00})(m_{11} - y_{11} + m_{10} - y_{10})} \right],$$

and the other from the corrected odds-ratio based on the remaining (unvalidated) data,

$$\log \widehat{OR}_R = \log \left[\frac{N_{11} N_{00}}{N_{10} N_{01}} \right], \tag{2.18}$$

where the N_{da}'s are the counts corrected for misclassification.

These corrected counts are the result of application of the matrix methods, where $E(\mathbf{N}) = \Phi^{-1}E(\mathbf{n})$, such that $\mathbf{N} = (N_{11}, N_{01}, N_{10}, N_{00})^T$, $\mathbf{n} = (n_{11} - m_{11}, n_{01} - m_{01}, n_{10} - m_{10}, n_{00} - m_{00})^T$ (the unvalidated observed data) and Φ is the matrix of classification probabilities,

$$\Phi = \begin{pmatrix} \theta_{11} & \theta_{10} & 0 & 0 \\ 1-\theta_{11} & 1-\theta_{10} & 0 & 0 \\ 0 & 0 & \theta_{01} & \theta_{00} \\ 0 & 0 & 1-\theta_{01} & 1-\theta_{00} \end{pmatrix}.$$

To estimate the corrected counts, the validation sample proportions are used, giving us the estimated classification matrix, denoted as $\mathbf{Q}$, and defined by,

$$\mathbf{Q} = \begin{pmatrix} \frac{y_{11}}{y_{11}+y_{10}} & \frac{m_{11}-y_{11}}{m_{11}-y_{11}+m_{10}-y_{10}} & 0 & 0 \\ \frac{y_{10}}{y_{11}+y_{10}} & \frac{m_{10}-y_{10}}{m_{11}-y_{11}+m_{10}-y_{10}} & 0 & 0 \\ 0 & 0 & \frac{y_{01}}{y_{01}+y_{00}} & \frac{m_{01}-y_{01}}{m_{01}-y_{01}+m_{00}-y_{00}} \\ 0 & 0 & \frac{y_{00}}{y_{01}+y_{00}} & \frac{m_{00}-y_{00}}{m_{01}-y_{01}+m_{00}-y_{00}} \end{pmatrix}.$$

Solving $\mathbf{N} = \mathbf{Q}^{-1}\mathbf{n}$ gives the N_{da} values needed to evaluate (2.18).

Greenland's estimator, defined in equation (11) of Greenland (2008) [13] is then a weighted average of the above two estimators, with respective weights,

$$\widehat{OR}_{IVW} = \left(\frac{1}{\hat{\sigma}^2(\log\widehat{OR}_R)} + \frac{1}{\hat{\sigma}^2(\log\widehat{OR}_V)}\right)^{-1}\left(\frac{\log\widehat{OR}_R}{\hat{\sigma}^2(\log\widehat{OR}_R)} + \frac{\log\widehat{OR}_V}{\hat{\sigma}^2(\log\widehat{OR}_V)}\right). \tag{2.19}$$

The asymptotic variance formula can be estimated by,

$$\hat{\sigma}^2(\log\widehat{OR}_{IVW}) = \left(\frac{1}{\hat{\sigma}^2(\log\widehat{OR}_R)} + \frac{1}{\hat{\sigma}^2(\log\widehat{OR}_V)}\right)^{-1}, \tag{2.20}$$

where $\hat{\sigma}^2(\log \widehat{OR}_V) = \frac{1}{y_{11}+y_{10}} + \frac{1}{y_{01}+y_{00}} + \frac{1}{m_{11}-y_{11}+m_{10}-y_{10}} + \frac{1}{m_{01}-y_{01}+m_{00}-y_{00}}$. The derivation of $\hat{\sigma}^2(\log \widehat{OR}_R)$ can be found in Greenland (1988) [12]. He introduced the formula as a special case of a general framework that was presented in the appendices of that paper. In Greenland (2008) [13], the variance formula was rewritten as the sum of two components denoted as $v_{UQ} = v_U + v_Q$, where,

$$v_Q = \sum_{d=0,1} \frac{(n_{d1} - m_{d1})(n_{d0} - m_{d0})(n_{d1} - m_{d1} + n_{d0} - m_{d0})}{(N_{d1} N_{d0} \det(\mathbf{Q}_d))^2}, \tag{2.21}$$

and

$$\begin{aligned} v_U \quad &= \sum_{d=0,1} \frac{\frac{1}{y_{d1}+y_{d0}} \left(\frac{y_{d1}}{y_{d1}+y_{d0}}\right) \left(1 - \frac{y_{d1}}{y_{d1}+y_{d0}}\right)}{\left(\frac{N_{d1}}{n_{d1}-m_{d1}}\right)^2 \det(\mathbf{Q}_d)^2} \\ &\quad + \frac{\frac{1}{m_{d1}-y_{d1}+m_{d0}-y_{d0}} \left(\frac{m_{d1}-y_{d1}}{m_{d1}-y_{d1}+m_{d0}-y_{d0}}\right) \left(1 - \frac{m_{d1}-y_{d1}}{m_{d1}-y_{d1}+m_{d0}-y_{d0}}\right)}{\left(\frac{N_{d0}}{n_{d0}-m_{d0}}\right)^2 \det(\mathbf{Q}_d)^2} \end{aligned} \tag{2.22}$$

where $\mathbf{Q}_d$ is the d^{th} diagonal block in $\mathbf{Q}$,

$$\mathbf{Q}_d = \begin{pmatrix} \frac{y_{d1}}{y_{d1}+y_{d0}} & \frac{m_{d1}-y_{d1}}{m_{d1}-y_{d1}+m_{d0}-y_{d0}} \\ \frac{y_{d0}}{y_{d1}+y_{d0}} & \frac{m_{d0}-y_{d0}}{m_{d1}-y_{d1}+m_{d0}-y_{d0}} \end{pmatrix}.$$

However, v_U was written incorrectly as,

$$v_U^* \quad = \sum_{d=0,1} \frac{y_{d1}\left(1 - \frac{y_{d1}}{y_{d1}+y_{d0}}\right)}{(N_{d1} \det(\mathbf{Q}_d))^2} + \frac{(m_{d1} - y_{d1})\left(1 - \frac{m_{d1}-y_{d1}}{m_{d1}-y_{d1}+m_{d0}-y_{d0}}\right)}{(N_{d0} \det(\mathbf{Q}_d))^2}, \tag{2.23}$$

which we believe was due to an algebraic error. We validated this claim by calculating v_u and v_u^* using the example data similarly provided in both papers. We note that the $v_u \neq v_u^*$; however, the same values are recorded in both papers and this is the value associated with v_u. Hence, we will apply the variance formula based on (2.21) and (2.22) in the following simulation study.

2.5 Comparative Simulation Study: IVW, 2S and 3S Approaches

As demonstrated in Section 2.3.1, the 3S approach outperforms the 2S approach for a range of P_1 values. We used three sets of parameter values to demonstrate this, and showed graphically that this range of values may have some relationship with the observed incidence rate. In this section, we investigate this behaviour in more detail, and conduct simulation studies comparing the IVW estimator with the 2S and 3S approaches.

To begin, we first assess the asymptotic properties and unbiasedness of the IVW estimator in a similar manner as in the previous section. The following are Tables 2.3.2b, 2.3.3b and 2.3.4b with the addition of the results for $\widehat{OR}_{IVW}$. Also, the 3S results are reported at the minimum only.

Table 2.5.1: Comparative simulation study results of odds-ratios estimated using 2S, 3S and IVW for parameter set 1 (OR = 0.4737).

Approach	Average Point Estimate	Average Standard Error	Sample Standard Deviation	CP
$\widetilde{OR}$	0.853	0.0677	0.0676	0.000
$\widehat{OR}_2$	0.476	0.1841	0.1864	0.953
$\widehat{OR}_3^{min}$	0.474	0.1450	0.1386	0.963
$\widehat{OR}_{IVW}$	0.491	0.1965	0.1906	0.955

Table 2.5.2: Comparative simulation study results of odds-ratios estimated using 2S, 3S and IVW for parameter set 2 (OR = 0.4444).

Approach	Average Point Estimate	Average Standard Error	Sample Standard Deviation	CP
$\widetilde{OR}$	0.49139	0.052217	0.050818	0.522
$\widehat{OR}_2$	0.45012	0.13127	0.12427	0.969
$\widehat{OR}_3^{min}$	0.45476	0.11148	0.1116	0.936
$\widehat{OR}_{IVW}$	0.45197	0.13358	0.12218	0.97

Table 2.5.3: Comparative simulation study results of odds-ratios estimated using 2S, 3S and IVW for parameter set 3 (OR = 1.7143).

Approach	Average Point Estimate	Average Standard Error	Sample Standard Deviation	CP
$\widetilde{OR}$	1.4671	0.043972	0.044676	0.062
$\widehat{OR}_2$	1.7436	0.10439	0.10839	0.938
$\widehat{OR}_3^{min}$	1.7357	0.097327	0.080904	0.995
$\widehat{OR}_{IVW}$	1.7359	0.10559	0.10183	0.968

Once again, due to the similarity in the results, only a small subset of results are presented here. Note that $\widehat{OR}_{IVW}$ appears unbiased and the estimates of standard error appear to approximate the sample standard deviation quite well. This approximation worsens at smaller values of m, although we have only reported those values associated with $n = 10,000$ and $m = 1000$. Thus, we will assume for the remainder of these studies that the IVW estimator produces unbiased estimates (as noted in Greenland (2008) [13]) and that its estimated asymptotic variance is a good approximation of the variability in the odds-ratio estimates.

Next, we can consider the estimated ARE. Based on the results in Tables 2.5.1-2.5.3 it appears that the 3S approach achieves the smallest estimated standard error for all three parameter sets. However, understanding how much better, is this statement generalizable and which values of P_1 to consider when planning a study are still important points to consider. Regarding the values of P_1, we will first formally prove the assertion that there exists a convex relationship between the 3S estimates of variance for the parameter of interest and P_1. (This point was demonstrated numerically in Figures 2.3.1 and 2.3.2 for parameter sets 1 and 3.) The property of convexity implies that a horizontal line (at the point of the 2S estimate of standard error for our purposes) with two intersection points will lie above the curve. This implies that if there are any values of P_1 that produce less variable 3S estimates, there will exist a range of P_1 values that produce more efficient estimates relative to the 2S approach. In other words, investigators will be able to select from a variety of validation sampling schemes to achieve a more efficient estimator.

Let γ be a p-dimensional parameter vector with corresponding multi-sample MLE, $\hat{\gamma}$, that follows the results of Theorem A.2.3 in Appendix A. By Theorem A.2.4 we have a consistent estimate of the asymptotic variance-covariance matrix in the form of equation (2.12). Next, let b be a unit vector designed to isolate the parameter of interest, eg. $b^T = (1, 0, 0, 0, 0, 0)$. Finally let A_1, A_2, A_3 be $k \times k$ positive definite matrices and $c \in [0, 1]$ and define,

$$f(c) = b^T(A_1 + cA_2 + (1 - c)A_3)^{-1}b$$

where $A_1 + cA_2 + (1 - c)A_3$ is also positive definite (since for $A_1 = n\bar{I}_O(\hat{\gamma})$, $A_1 = m\bar{I}_{V_1}(\hat{\gamma})$, $A_3 = m\bar{I}_{V_0}(\hat{\gamma})$ the sum is simply $I_3(\hat{\gamma})$).

Theorem 2.2.1: $f(c)$ is a convex function. Let $\theta, x, y \in [0, 1]$. Then,

Proof

$$\begin{aligned}
f(\theta x + (1-\theta)y) \quad &= \quad b^T\left(A_1 + [(\theta x + (1-\theta)y)A_2 + (1 - \theta x - (1-\theta)y)A_3]\right)^{-1} b \\
&= b^T(\theta[A_1 + xA_2 + A_3(1-x)] + (1-\theta)[A_1 + yA_2 + A_3(1-y)])^{-1}b \\
&\overset{*}{\leq} \theta b^T[A_1 + xA_2 - xA_3 + A_3]^{-1}b + (1-\theta)b^T[A_1 + yA_2 - yA_3 + A_3]^{-1}b \\
&= \theta b^T[A_1 + xA_2 + (1-x)A_3]^{-1}b + (1-\theta)b^T[A_1 + yA_2 + (1-y)A_3]^{-1}b \\
&= \theta f(x) + (1-\theta)f(y),
\end{aligned}$$

where (*) is due to the result,

$$\left(\sum_i \omega_i A_i\right)^{-1} \preceq \sum_i \omega_i A_i^{-1},$$

where $0 \leq \omega_i \leq 1$ for all i and $\sum_i \omega_i = 1$ (see Kagan et al. (2010) [18]). Also, $A_1 + cA_2 + (1 - c)A_3$ will be positive definite, for both $c = x, y$, by assumption.

Thus, the 3S approach's estimates of variance associated with the estimator for the parameter of interest will have a convex relationship with P_1. Recall that there appeared to be a value of P_1 at which the 3S approach represents the more efficient approach, as compared to the IVW or 2S approaches. This value seems to be related to the observed incidence rate, as can be observed in the plots in Section 2.3.1. Thus, we will conduct more extensive simulation studies to investigate this further. For Tables 2.5.4 - 2.5.16 we present a number of results designed to numerically investigate these transition points, and approximate the resulting interval in which the 3S approach may outperform the 2S and IVW approaches.

To begin, we define the interval of P_1 values in which the 3S approach performs better than the 2S and IVW approaches as (L_2, U_2) and (L_{IVW}, U_{IVW}) respectively. For example, in Table 2.5.4, the third entry is based on $\pi_{1|1} = 0.2$, $\pi_{1|0} = 0.2$, $\theta_{11} = 0.95$, $\theta_{01} = 0.95$, $\theta_{10} = 0.05$, $\theta_{00} = 0.05$ with equal values for $n_{d.}$, $d = 0, 1$. This yields an expected incidence rate of 0.23 and produces the intervals $(L_2 = 0.22, U_2 = 0.73)$ for the 2S approach and $(L_{IVW} = 0.22, U_{IVW} = 0.73)$ for the IVW approach. This can be interpreted as the approximate range of P_1 values for which the 3S approach will produce lower estimates of standard error than the 2S and IVW approaches, respectively. However, as this is only the upper and lower bound, we report the minimum estimated standard error for the 3S approach and the estimated standard errors for the other approaches as well. For our example, this gives $\hat{\sigma}^2(\log \widehat{OR}_3) = 0.0066$ at the minimal P_1, $\hat{\sigma}^2(\log \widehat{OR}_2) = 0.0086$ and $\hat{\sigma}^2(\log \widehat{OR}_{IVW}) = 0.0087$.

Note, that there are a number of practical issues that could cause computational problems. For instance, for very small (or large) values of P_1, we may have difficulty sampling enough observations to properly estimate the parameters of interest. This could lead to singularity of the information matrix or the matrix of probabilities in the IVW method. Thus, these iterations will simply be ignored as this would be an issue in practical applications as well. Finally, the chosen simulation parameter values alter the target odds-ratio values as well as the θ-parameters. We have first chosen to consider three odds-ratio values of (0.5,1 and 2) with increasing rates of misclassification in a non-differential manner as seen in Tables 2.5.4-2.5.6. Next, we select target parameters designed to produce increasing expected incidence rates of category 1 observations in the original data. These results are displayed in Table 2.5.7. Finally, using the same odds-ratio values as in Tables 2.5.4-2.5.6, we consider a variety of θ-parameters designed to consider differing patterns of misclassification and present those results in Tables 2.5.8-2.5.10. Once again, we report the results for $n = 10,000$ and $m = 1,000$ and initially consider equal exposure group sizes.

Table 2.5.4: Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 1$, $n = 10,000$, $m = 1,000$.

	Target Parameters						2S Interval		IVW Interval		MC Estimates of Variance		
$\frac{E(n_{.1})}{n}$	$\pi_{1\|1}$	$\pi_{1\|0}$	θ_{11}	θ_{01}	θ_{10}	θ_{00}	Lower	Upper	Lower	Upper	2S	IVW	3S min
0.206	0.2	0.2	0.99	0.99	0.01	0.01	0.2	0.57	0.2	0.57	0.004	0.004	0.0037
0.218	0.2	0.2	0.97	0.97	0.03	0.03	0.21	0.69	0.21	0.69	0.0064	0.0065	0.0051
0.23	0.2	0.2	0.95	0.95	0.05	0.05	0.22	0.73	0.22	0.73	0.0086	0.0087	0.0066
0.248	0.2	0.2	0.92	0.92	0.08	0.08	0.24	0.74	0.24	0.74	0.0115	0.0116	0.0088
0.26	0.2	0.2	0.9	0.9	0.1	0.1	0.26	0.75	0.24	0.75	0.0132	0.0134	0.0103
0.278	0.2	0.2	0.87	0.87	0.13	0.13	0.28	0.72	0.27	0.73	0.0151	0.0154	0.0123
0.29	0.2	0.2	0.85	0.85	0.15	0.15	0.29	0.62	0.27	0.73	0.0163	0.0168	0.0135

Table 2.5.5: Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 0.5$, $n = 10,000$, $m = 1,000$.

	Target Parameters						2S Interval		IVW Interval		MC Estimates of Variance		
$\frac{E(n_{.1})}{n}$	$\pi_{1\|1}$	$\pi_{1\|0}$	θ_{11}	θ_{01}	θ_{10}	θ_{00}	Lower	Upper	Lower	Upper	2S	IVW	3S min
0.271	0.2	0.333	0.99	0.99	0.01	0.01	0.27	0.57	0.26	0.57	0.0033	0.0033	0.0031
0.281	0.2	0.333	0.97	0.97	0.03	0.03	0.27	0.68	0.27	0.68	0.0052	0.0052	0.0045
0.29	0.2	0.333	0.95	0.95	0.05	0.05	0.29	0.71	0.28	0.71	0.0069	0.007	0.0058
0.304	0.2	0.333	0.92	0.92	0.08	0.08	0.28	0.72	0.28	0.72	0.0094	0.0095	0.0076
0.313	0.2	0.333	0.9	0.9	0.1	0.1	0.22	0.64	0.22	0.64	0.0109	0.0108	0.0087
0.327	0.2	0.333	0.87	0.87	0.13	0.13	0.32	0.69	0.31	0.7	0.0125	0.0127	0.0109
0.337	0.2	0.333	0.85	0.85	0.15	0.15	0.32	0.67	0.32	0.68	0.0136	0.0137	0.012

Table 2.5.6: Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 2$, $n = 10,000$, $m = 1,000$.

	Target Parameters						2S Interval		IVW Interval		MC Estimates of Variance		
$\frac{E(n_{.1})}{n}$	$\pi_{1\|1}$	$\pi_{1\|0}$	θ_{11}	θ_{01}	θ_{10}	θ_{00}	Lower	Upper	Lower	Upper	2S	IVW	3S min
0.328	0.4	0.25	0.99	0.99	0.01	0.01	0.31	0.56	0.31	0.56	0.0028	0.0028	0.0027
0.336	0.4	0.25	0.97	0.97	0.03	0.03	0.33	0.64	0.33	0.64	0.0042	0.0043	0.0039
0.342	0.4	0.25	0.95	0.95	0.05	0.05	0.34	0.66	0.33	0.66	0.0057	0.0057	0.0052
0.353	0.4	0.25	0.92	0.92	0.08	0.08	0.35	0.66	0.34	0.66	0.0076	0.0077	0.0069
0.36	0.4	0.25	0.9	0.9	0.1	0.1	0.37	0.38	0.35	0.36	0.0086	0.0088	0.008
0.37	0.4	0.25	0.87	0.87	0.13	0.13	0.37	0.64	0.36	0.66	0.0104	0.0106	0.0097
0.378	0.4	0.25	0.85	0.85	0.15	0.15	0.37	0.63	0.37	0.64	0.0114	0.0115	0.0106

Table 2.5.7: Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for increasing values of $\frac{n_{.1}}{n}$, $n = 10,000$, $m = 1,000$.

	Target Parameters						2S Interval		IVW Interval		MC Estimates of Variance		
$\frac{E(n_{.1})}{n}$	$\pi_{1\|1}$	$\pi_{1\|0}$	θ_{11}	θ_{01}	θ_{10}	θ_{00}	Lower	Upper	Lower	Upper	2S	IVW	3S min
0.097	0.05	0.1	0.85	0.95	0.05	0.01	-	0.69	-	0.72	0.0346	0.0386	0.0206
0.13	0.1	0.1	0.85	0.85	0.05	0.05	0.13	0.67	0.12	0.67	0.0243	0.0249	0.0171
0.157	0.2	0.1	0.85	0.95	0.05	0.01	0.15	0.54	0.13	0.58	0.0109	0.0116	0.0093
0.159	0.1	0.2	0.95	0.85	0.05	0.01	0.15	0.66	0.15	0.67	0.0142	0.0146	0.0105
0.246	0.3	0.2	0.85	0.95	0.01	0.05	0.25	0.49	0.25	0.49	0.0072	0.0072	0.0067
0.276	0.2	0.3	0.95	0.95	0.01	0.1	0.26	0.67	0.23	0.7	0.0065	0.0068	0.0055
0.3	0.3	0.2	0.9	0.9	0.1	0.1	0.3	0.71	0.28	0.74	0.0111	0.0113	0.0091
0.36	0.4	0.25	0.9	0.9	0.1	0.1	0.37	0.38	0.35	0.64	0.0086	0.0088	0.008

Table 2.5.8: Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 1$, $n = 10,000$, $m = 1,000$.

	Target Parameters						2S Interval		IVW Interval		MC Estimates of Variance		
$\frac{E(n_{.1})}{n}$	$\pi_{1\|1}$	$\pi_{1\|0}$	θ_{11}	θ_{01}	θ_{10}	θ_{00}	Lower	Upper	Lower	Upper	2S	IVW	3S min
0.218	0.2	0.2	0.99	0.95	0.01	0.05	0.21	0.68	0.19	0.7	0.0063	0.0066	0.0051
0.214	0.2	0.2	0.93	0.97	0.05	0.01	0.21	0.64	0.19	0.66	0.0068	0.0071	0.0057
0.204	0.2	0.2	0.85	0.95	0.01	0.05	0.19	0.5	0.19	0.5	0.0081	0.0082	0.0074
0.206	0.2	0.2	0.9	0.8	0.08	0.01	0.2	0.49	0.19	0.51	0.0105	0.0107	0.0096
0.32	0.2	0.2	0.9	0.9	0.15	0.2	0.32	0.79	0.31	-	0.0161	0.0166	0.0123
0.28	0.2	0.2	0.83	0.97	0.15	0.1	0.28	0.75	0.25	0.78	0.0143	0.015	0.011
0.23	0.2	0.2	0.85	0.85	0.1	0.05	0.24	0.57	0.24	0.59	0.0124	0.0126	0.0109

Table 2.5.9: Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 0.5$, $n = 10,000$, $m = 1,000$.

	Target Parameters						2S Interval		IVW Interval		MC Estimates of Variance		
$\frac{E(n_{.1})}{n}$	$\pi_{1\|1}$	$\pi_{1\|0}$	θ_{11}	θ_{01}	θ_{10}	θ_{00}	Lower	Upper	Lower	Upper	2S	IVW	3S min
0.278	0.2	0.333	0.99	0.95	0.01	0.05	0.28	0.61	0.25	0.64	0.0046	0.0047	0.0042
0.278	0.2	0.333	0.93	0.97	0.05	0.01	0.28	0.63	0.24	0.67	0.006	0.0063	0.0054
0.264	0.2	0.333	0.85	0.95	0.01	0.05	0.27	0.36	0.24	0.41	0.0064	0.0066	0.0063
0.259	0.2	0.333	0.9	0.8	0.08	0.01	0.28	0.46	0.28	0.44	0.0092	0.0092	0.0085
0.367	0.2	0.333	0.9	0.9	0.15	0.2	0.38	0.75	0.36	0.37	0.0128	0.0131	0.0108
0.338	0.2	0.333	0.83	0.97	0.15	0.1	0.34	0.72	0.28	0.76	0.0117	0.0128	0.01
0.283	0.2	0.333	0.85	0.85	0.1	0.05	0.29	0.54	0.27	0.57	0.0107	0.0111	0.0097

Table 2.5.10: Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 2$, $n = 10,000$, $m = 1,000$.

	Target Parameters						2S Interval		IVW Interval		MC Estimates of Variance		
$\frac{E(n_{.1})}{n}$	$\pi_{1\|1}$	$\pi_{1\|0}$	θ_{11}	θ_{01}	θ_{10}	θ_{00}	Lower	Upper	Lower	Upper	2S	IVW	3S min
0.338	0.4	0.25	0.99	0.95	0.01	0.05	0.34	0.67	0.3	0.7	0.0045	0.0048	0.0041
0.326	0.4	0.25	0.93	0.97	0.05	0.01	0.32	0.5	0.28	0.56	0.0044	0.0045	0.0043
0.31	0.4	0.25	0.85	0.95	0.01	0.05	0.3	0.45	0.28	0.46	0.0061	0.0061	0.0059
0.308	0.4	0.25	0.9	0.8	0.08	0.01	0.24	0.25	0.2	0.21	0.0072	0.0074	0.007
0.412	0.4	0.25	0.9	0.9	0.15	0.2	0.41	0.74	0.38	0.76	0.0112	0.0116	0.0099
0.37	0.4	0.25	0.83	0.97	0.15	0.1	0.36	0.69	0.33	0.71	0.0093	0.0097	0.0084
0.325	0.4	0.25	0.85	0.85	0.1	0.05	0.31	0.5	0.29	0.51	0.0087	0.0088	0.0084

In total, we have presented the results of 50 different simulation studies. For all studies, the 3S approach appears to outperform the 2S approach and the IVW method. In Table 2.5.7, the first row contains '-' for the lower bounds of the interval with respect to both methods. This indicates that the 3S approach is producing smaller estimates of variance at the low end of considered P_1 values, $P_1 = 0.1$. The same can be said about Table 2.5.8, row 5 in which the upper bound does not have a value, however, the upper end of the range is at $P_1 = 0.8$ for these simulations and the transition point likely lies above this value.

Next, we discuss the relationship of the intervals with the observed incidence rate, $\frac{n_{.1}}{n}$. Some interesting trends are visible in the results displayed in Tables 2.5.4-2.5.10. In most cases, the lower bound for both intervals appear to exhibit an underlying pattern that lines up with the observed incidence rates. Since we did not consider incidence rates above 50%, we ran two sets of simulations to investigate this trend when the incidence rate of category 1 observations is large. To do so, we have chosen a value of the odds-ratio of 2, but considered π-parameters that would cause the category 1 incidence rate to be greater than 0.5. Clearly, the lower bounds are not close to these rates, however, the upper bounds now appear to approximate the incidence rates for most of the simulations. Note that these transition points are based on Monte Carlo estimates and as such, these values are subject to sampling variability. Since this interval represents the boundaries of P_1 in which the 3S approach will outperform the 2S and IVW approaches, it appears that at least one of the values is possibly related to the incidence rates.

Table 2.5.11: Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for incidence rates greater than 0.5, $OR = 2$, $n = 10,000$, $m = 1,000$.

	Target Parameters						2S Interval		IVW Interval		MC Estimates of Variance		
$\frac{E(n_{.1})}{n}$	$\pi_{1\|1}$	$\pi_{1\|0}$	θ_{11}	θ_{01}	θ_{10}	θ_{00}	Lower	Upper	Lower	Upper	2S	IVW	3S min
0.729	0.8	0.667	0.99	0.99	0.01	0.01	0.46	0.55	0.46	0.55	0.0033	0.0033	0.003
0.719	0.8	0.667	0.97	0.97	0.03	0.03	0.31	0.71	0.31	0.72	0.0052	0.0052	0.0045
0.71	0.8	0.667	0.95	0.95	0.05	0.05	0.25	0.59	0.25	0.59	0.0062	0.0063	0.0052
0.696	0.8	0.667	0.92	0.92	0.08	0.08	0.28	0.7	0.3	0.31	0.0095	0.0093	0.0076
0.687	0.8	0.667	0.9	0.9	0.1	0.1	0.25	0.62	0.25	0.62	0.0105	0.0105	0.0089
0.673	0.8	0.667	0.87	0.87	0.13	0.13	0.31	0.66	0.3	0.68	0.0124	0.0125	0.0109
0.663	0.8	0.667	0.85	0.85	0.15	0.15	0.31	0.66	0.3	0.67	0.0135	0.0137	0.012

Table 2.5.12: Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for incidence rates greater than 0.5, $OR = 2$, $n = 10,000$, $m = 1,000$.

	Target Parameters						2S Interval		IVW Interval		MC Estimates of Variance		
$\frac{E(n_{.1})}{n}$	$\pi_{1\|1}$	$\pi_{1\|0}$	θ_{11}	θ_{01}	θ_{10}	θ_{00}	Lower	Upper	Lower	Upper	2S	IVW	3S min
0.722	0.8	0.667	0.99	0.95	0.01	0.05	0.38	0.71	0.36	0.74	0.0046	0.0047	0.0042
0.702	0.8	0.667	0.93	0.97	0.05	0.01	0.23	0.7	0.22	0.72	0.0068	0.0071	0.0054
0.666	0.8	0.667	0.85	0.95	0.01	0.05	0.17	0.67	0.16	0.69	0.0094	0.0099	0.0068
0.636	0.8	0.667	0.9	0.8	0.08	0.01	0.17	0.64	0.17	0.64	0.011	0.0111	0.0083
0.708	0.8	0.667	0.9	0.9	0.15	0.2	0.4	0.71	0.39	0.72	0.0123	0.0124	0.0112
0.687	0.8	0.667	0.83	0.97	0.15	0.1	0.34	0.69	0.26	0.75	0.0114	0.0128	0.0101
0.642	0.8	0.667	0.85	0.85	0.1	0.05	0.2	0.58	0.2	0.58	0.0121	0.0121	0.0096

Although altering the sample size will naturally change the estimates of variance the relationship we observed with the incidence rates remains the same. In

Tables 2.5.13-2.5.14 we display two of the results for $n = 100,000$ and $m = 1,000$ with an $OR = 2$ to consider a much larger sample size.

Table 2.5.13: Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 2$, $n = 100,000$, $m = 1,000$.

	Target Parameters						2S Interval		IVW Interval		MC Estimates of Variance		
$\frac{E(n_{.1})}{n}$	$\pi_{1\|1}$	$\pi_{1\|0}$	θ_{11}	θ_{01}	θ_{10}	θ_{00}	Lower	Upper	Lower	Upper	2S	IVW	3S min
0.328	0.400	0.250	0.99	0.99	0.01	0.01	0.32	0.53	0.32	0.54	0.0012	0.0012	0.0011
0.336	0.400	0.250	0.97	0.97	0.03	0.03	0.33	0.65	0.33	0.65	0.0028	0.0028	0.0025
0.342	0.400	0.250	0.95	0.95	0.05	0.05	0.35	0.66	0.34	0.67	0.0043	0.0044	0.0039
0.353	0.400	0.250	0.92	0.92	0.08	0.08	0.36	0.66	0.34	0.67	0.0065	0.0066	0.0058
0.360	0.400	0.250	0.90	0.90	0.10	0.10	0.36	0.66	0.35	0.66	0.0078	0.0079	0.0071
0.370	0.400	0.250	0.87	0.87	0.13	0.13	0.37	0.64	0.36	0.65	0.0096	0.0097	0.0089
0.378	0.400	0.250	0.85	0.85	0.15	0.15	0.38	0.63	0.36	0.65	0.0107	0.0108	0.0099

Table 2.5.14: Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 2$, $n = 100,000$, $m = 1,000$.

	Target Parameters						2S Interval		IVW Interval		MC Estimates of Variance		
$\frac{E(n_{.1})}{n}$	$\pi_{1\|1}$	$\pi_{1\|0}$	θ_{11}	θ_{01}	θ_{10}	θ_{00}	Lower	Upper	Lower	Upper	2S	IVW	3S min
0.338	0.400	0.250	0.99	0.95	0.01	0.05	0.34	0.66	0.32	0.72	0.0031	0.0032	0.0026
0.326	0.400	0.250	0.93	0.97	0.05	0.01	-	-	0.32	0.51	0.0028	0.0029	0.0028
0.310	0.400	0.250	0.85	0.95	0.01	0.05	0.34	0.42	0.33	0.42	0.0047	0.0047	0.0046
0.308	0.400	0.250	0.90	0.80	0.08	0.01	-	-	0.36	0.36	0.0058	0.0059	0.0059
0.412	0.400	0.250	0.90	0.90	0.15	0.20	0.42	0.74	0.39	0.75	0.0104	0.0107	0.0091
0.370	0.400	0.250	0.83	0.97	0.15	0.10	0.38	0.68	0.35	0.71	0.0083	0.0087	0.0075
0.325	0.400	0.250	0.85	0.85	0.10	0.05	0.33	0.42	0.33	0.48	0.0075	0.0076	0.0073

In Table 2.5.14, two studies demonstrate that the 2S approach outperforms the 3S approach since no interval is produced. In other words, the 2S approach produces smaller estimates of variance than all of the 3S estimates. However, considering the estimated variances, $\hat{\sigma}^2(\log \widehat{OR}_2) = 0.0028$ and $\hat{\sigma}^2(\log \widehat{OR}_3) = 0.0028$ for the results in row 2 and $\hat{\sigma}^2(\log \widehat{OR}_2) = 0.0058$ and $\hat{\sigma}^2(\log \widehat{OR}_3) = 0.0059$ in row 4, lead us to conclude that these are also approximately equivalent. Next, consider row 3 in which the estimates of variance are also quite close together however a range of values is produced of $(L_2 = 0.34, U_2 = 0.42)$ for the 2S approach. This situation demonstrates the desirability of the 3S approach to choose a validation sample with a category 1 sample size ranging between $m_{.1} = 0.34 \times m$ and $m_{.1} = 0.42 \times m$, while attaining similar precision as the 2S approach with

$m_{\cdot 1} \approx 0.31 \times m$. For those scenarios in which the 2S and 3S approach perform in a similar manner, it appears that there are fairly wide intervals of P_1 values for very small changes in the estimated variance. In other words, utilizing the 3S approach will allow us to choose from a wide range of validation sample schemes. However, it seems possible that allowing for slightly higher variance will increase the range of acceptable P_1 values appreciably. For example, if minimizing the category 1 validation sample size is more important than a slight loss in precision, perhaps selecting a value of $P_1 < \frac{n_{\cdot 1}}{n}$ may be ideal. We will investigate this graphically in the following plots that demonstrate the estimates of variance across a range of the P_1 values for two examples. The first is the previously discussed example (the third row of Table 2.5.14) and the second is the fourth row in the same table. Note that the first example has a set of values in which the 3S approach is the best, while in the second example, the 2S approach is better. The solid line on the plots in Figures 2.5.1 and 2.5.2 denotes the Monte Carlo estimate of the 2S standard deviation and the dotted line represents 0.001 above that estimate.

Figure 2.5.1: Plot of Monte Carlo estimates of $\max(\hat{\sigma}(\hat{\pi}_{1|1}), \hat{\sigma}(\hat{\pi}_{1|0}))$ versus P_1 for example 1 (third row of Table 2.5.14).

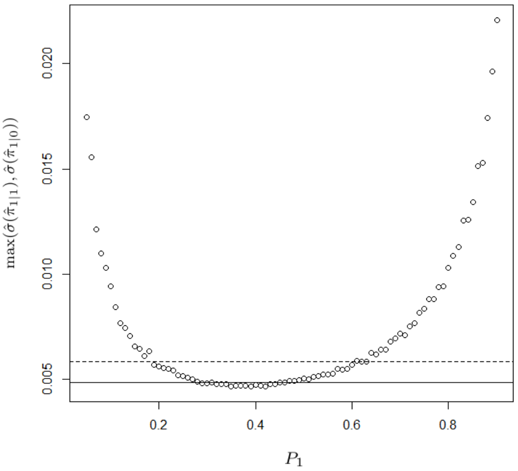
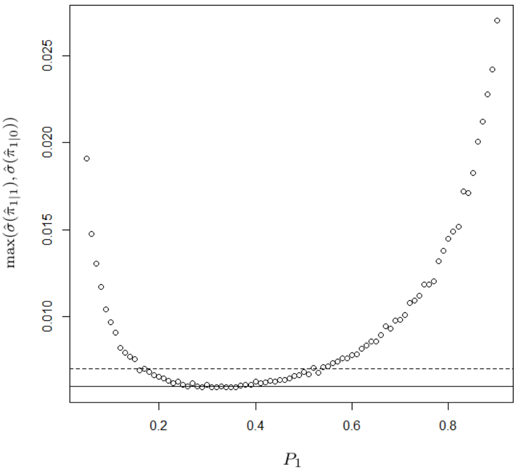


Figure 2.5.2: Plot of Monte Carlo estimates of $\max(\hat{\sigma}(\hat{\pi}_{1|1}), \hat{\sigma}(\hat{\pi}_{1|0}))$ versus P_1 for example 2 (fourth row of Table 2.5.14).



It is evident that allowing for such a small difference will increase the width of the intervals substantially. In the first plot, it appears that the new interval allows us to select a value of $m_{.1}$ between approximately $m_{.1} = 0.19 \times m$ and $m_{.1} = 0.61 \times m$, while the original values were $L_2 = 0.34$ and $U_2 = 0.42$. The second example has produced an interval of length zero, (L_2, U_2), since the estimated variance for the 2S approach is smaller for all P_1. However, the plot demonstrates that there is a range of values in which the estimates of variance are extremely close, allowing for the selection of $m_{.1}$ in the range $m_{.1} = 0.25 \times m$ to $m_{.1} = 0.38 \times m$. Note that this interval contains the incidence rate, $\frac{n_{.1}}{n} = 0.3$. Increasing our tolerance for variability in the estimates up slightly to the dotted line increases the range to approximately $m_{.1} = 0.18 \times m$ to $m_{.1} = 0.52 \times m$.

Finally, alteration of the exposure group sizes for a fixed n is another question of interest since the groups size will alter the incidence rates as well as the estimates of variance. To investigate this, we ran a set of simulations with $n = 10,000$, $m = 1,000$ with the exposure groups set to be 80% for $D = 1$ and 20% for $D = 0$ followed by 20% and 80%. We conducted 1000 simulation iterations and list the simulation target parameters in results tables. Again, due to similarity of these results, we will only present two sets of output for $OR = 1$. The estimated variances appear to increase (relative to the case of equal exposure group sizes) as the exposure group sizes deviates from 50% in each group. However, it is important to note that the patterns described in the previous simulations with regards to the ARE are similar, implying that altering the exposure group sizes does not influence those relationships.

Table 2.5.15: Estimated range of P_1 values for which $\hat{\sigma}^2(\log\widehat{OR}_2)$ or $\hat{\sigma}^2(\log\widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log\widehat{OR}_3)$ for $OR = 1$, $n_{1.} = 8000$, $n_{0.} = 2000$, $m = 1,000$.

	Target Parameters						2S Interval		IVW Interval		MC Estimates of Variance		
$\frac{E(n_{.1})}{n}$	$\pi_{1\|1}$	$\pi_{1\|0}$	θ_{11}	θ_{01}	θ_{10}	θ_{00}	Lower	Upper	Lower	Upper	2S	IVW	3S min
0.225	0.200	0.200	0.99	0.95	0.01	0.05	0.25	0.52	0.22	0.57	0.0082	0.0087	0.0075
0.207	0.200	0.200	0.93	0.97	0.05	0.01	0.21	0.62	0.19	0.65	0.0129	0.0136	0.0106
0.220	0.200	0.200	0.85	0.95	0.01	0.05	0.23	0.38	0.20	0.44	0.0125	0.0130	0.0121
0.183	0.200	0.200	0.90	0.80	0.08	0.01	0.19	0.60	0.18	0.63	0.0179	0.0185	0.0144
0.332	0.200	0.200	0.90	0.90	0.15	0.20	0.33	0.76	0.32	0.77	0.0251	0.0256	0.0198
0.276	0.200	0.200	0.83	0.97	0.15	0.10	0.27	0.71	0.26	0.73	0.0252	0.0263	0.0200
0.218	0.200	0.200	0.85	0.85	0.10	0.05	0.22	0.64	0.21	0.66	0.0217	0.0223	0.0177

Table 2.5.16: Estimated range of P_1 values for which $\hat{\sigma}^2(\log \widehat{OR}_2)$ or $\hat{\sigma}^2(\log \widehat{OR}_{IVW})$ exceeds $\hat{\sigma}^2(\log \widehat{OR}_3)$ for $OR = 1$, $n_{1.} = 2000$, $n_{0.} = 8000$, $m = 1,000$.

	Target Parameters						2S Interval		IVW Interval		MC Estimates of Variance		
$\frac{E(n_{.1})}{n}$	$\pi_{1\|1}$	$\pi_{1\|0}$	θ_{11}	θ_{01}	θ_{10}	θ_{00}	Lower	Upper	Lower	Upper	2S	IVW	3S min
0.211	0.200	0.200	0.99	0.95	0.01	0.05	0.23	0.64	0.21	0.66	0.0120	0.0126	0.0097
0.221	0.200	0.200	0.93	0.97	0.05	0.01	0.25	0.50	0.21	0.55	0.0088	0.0095	0.0082
0.188	0.200	0.200	0.85	0.95	0.01	0.05	0.20	0.59	0.19	0.60	0.0131	0.0135	0.0110
0.229	0.200	0.200	0.90	0.80	0.08	0.01	0.26	0.31	0.20	0.42	0.0150	0.0157	0.0149
0.308	0.200	0.200	0.90	0.90	0.15	0.20	0.32	0.77	0.31	0.77	0.0265	0.0269	0.0201
0.284	0.200	0.200	0.83	0.97	0.15	0.10	0.30	0.74	0.27	0.75	0.0197	0.0208	0.0151
0.242	0.200	0.200	0.85	0.85	0.10	0.05	0.25	0.57	0.22	0.60	0.0184	0.0191	0.0164

These simulation studies have provided a perspective on the relative performance of these estimators across a number of realistic scenarios. We have shown consistently that drawing a validation sample conditional on the original observed realization of the outcome classifier can, under the right sampling scheme, produce estimates that are at least as efficient as the approaches that gather validation data via random sampling. Further, between the random sampling approaches, it appears that the IVW approach consistently produce larger estimates of variance as compared to the 2S approach. This observation may not imply that they are statistically different across all simulations, but it does imply that relative to these results, the 2S approach will be at least as efficient as the IVW approach. Thus, we will restrict the remainder of this discussion to the practical application of the 2S and 3S approaches.

As noted throughout this section, there appears to be a range of P_1 values in which the 3S approach produces more efficient estimates of the MCB-adjusted odds-ratio. In practice, however, the details of this range of values are generally unknown. We have demonstrated that the observed incidence rate seems to be related to one of the boundaries of this interval, the lower if $\frac{n_{.1}}{n} < 0.5$ and the upper otherwise. The value of the other bound is harder to determine. Attempting to solve the equation $\hat{\sigma}^2(\log \widehat{OR}_2) = \hat{\sigma}^2(\log \widehat{OR}_3)$ for P_1 would require knowledge of the individual components of variance, $\bar{I}_s(\hat{\gamma})$ in equations (2.15) and (2.17), for each sample. However, it does appear that the other bound is located above 50% in the majority of simulations, and vice versa when the observed incidence rate of category 1 observations in the original sample exceeds 50%. The simulations in which this is not the case appear to be the situations in which the methods are equally efficient.

This makes intuitive sense due to the convex relationship between $\hat{\sigma}^2(\log \widehat{OR}_3)$ and P_1. Since a bound of the interval corresponds to the value of P_1 such that $\hat{\sigma}^2(\log \widehat{OR}_2) \approx \hat{\sigma}^2(\log \widehat{OR}_3)$, if $\min_{P_1}[\hat{\sigma}^2(\log \widehat{OR}_3)]$ is close to $\hat{\sigma}^2(\log \widehat{OR}_2)$, then the width of the interval will be small. Thus, if $\hat{\sigma}^2(\log \widehat{OR}_2)$ is larger than $\hat{\sigma}^2(\log \widehat{OR}_3)$ the range of P_1 values increases, and the minimum shifts past the observed incidence rate, ie. $\frac{n_{.1}}{n} < \min_{P_1}[\hat{\sigma}^2(\log \widehat{OR}_3)]$ for $\frac{n_{.1}}{n} < 0.5$. This implies that the interval length will likely increase as the estimates of variance get farther apart. Further, for the scenarios in which the estimates of variance are extremely close, we have demonstrated numerically that slightly relaxing the desire for minimal variance will increase the interval size, making the upper transition point increase in value.

The results of this analysis suggest a rough and ready method for selecting values of P_1 that allows investigators to attain either a more efficient estimator, or an estimator that is approximately as efficient but allows for more freedom in the selection of a validation sampling scheme. Prior to drawing a validation sample, we already have the original data and access to the incidence rate of category 1 observations. Thus, based on the previous discussion, a reasonable choice for P_1 is a value in the interval, $[\frac{n_{.1}}{n}, 0.5]$ ($[0.5, \frac{n_{.1}}{n}]$ if $\frac{n_{.1}}{n} > 0.5$). Based on our simulation results this suggestion would only fail for those situations in which both the 2S and 3S approach at $\min_{P_1}[\hat{\sigma}^2(\log \widehat{OR}_3)]$ perform similarly, since the interval length will decrease to zero. However, we have shown that small changes in estimated variance increased the range of P_1 considerably. This implies that selecting a value of P_1 that is not $\min_{P_1}[\hat{\sigma}^2(\log \widehat{OR}_3)]$ but is close to the bounds of the interval will likely only slightly increase (or decrease) the estimated variance relative to the 2S approach. Therefore, if our suggested range is not entirely accurate, the additional variability incurred will likely be small.

This suggestion serves only as a guide that is likely to produce an estimator that is at least as (and often more) efficient than the 2S approach, with the added benefit of giving an investigator the option to select a particular validation sampling scheme from a wider array of possible values of $m_{.1}$. However, the ideal value of P_1 is not attainable without some prior additional information, and as such there is an interest in further investigation through Monte Carlo simulation. The creation of a validation sample size determination algorithm that considers P_1 as well as bounds the standard error is a natural extension of this work that would produce a realistic suggestion in practice. Ideally, the ability to select the value of P_1 that minimizes $\hat{\sigma}^2(\log \widehat{OR}_3)$ will allow investigators to design optimal validation sampling schemes that minimize cost and/or variability.

2.6 Conclusions and Discussion

We have demonstrated through Monte Carlo simulation that the use of internal validation data to adjust for binary outcome misclassification bias can be accomplished using estimators constructed on the 2S or 3S approaches. We have also demonstrated numerically that the use of random sampling to attain this additional validation information will produce estimates that are at best as asymptotically efficient as those derived from a validation sample drawn conditionally on the possibly misclassified observed outcome status for ideal values of P_1. Further we have shown that the 2S approach will restrict the categorical make-up of the resulting validation sample to an interval about the observed incidence rate of category 1 observations in the original sample. Thus, the ideal and most practical solution is to use the conditional sampling maximum likelihood approach outlined in Section 2.2.

However, as mentioned in Section 2.5, there is a wide range of P_1 values that can be used to attain estimators at least as efficient as in the 2S approach. This range is not easy to predict, as it is a function of the asymptotic variance of the MCB-adjusted estimator, which in turn is estimated by a function of the resulting validation sample counts. In the next chapter, we will develop a numerical approach to approximate these sample counts which will allow for the development of a validation sample size determination algorithm. Ideally, investigators would design their validation sampling schemes based on a reasonable approximation of the value of $\min_{P_1}[\hat{\sigma}^2(\log \widehat{OR}_3)]$. Further, an algorithm of this type could allow investigators to consider possible trade-offs between sampling costs and resulting efficiency in estimation, which could help in selecting the ideal sampling approach.

Finally, this paper has been focused on the odds-ratio, however, these results are applicable to other contingency table methods of estimation such as the relative risk. All the necessary theoretical framework is based on estimation of binomial success probabilities when the observed binary outcome is subject to misclassification. A further extension of this work can incorporate covariate information via logistically modelling the π and θ-parameters. Lyles (2011) [25] discussed a random validation sampling approach to correct for MCB in logistic regression using internal validation data. Thus, there is a need to extend the conditional validation sampling approach described here to these types of models.

Chapter 3

Monte Carlo Sample Size Determination for Unbiased Estimation with Validation Data in the Presence of Binary Outcome Misclassification

3.1 Introduction

In the previous section, we investigated the efficiency of the misclassification bias (MCB) adjusted maximum likelihood estimators (MLEs) derived using additional internal validation information; drawn using a random sampling approach (the 2S approach) or a conditional sampling approach (the 3S approach) with different sampling schemes. We were able to show numerically, that under the 3S approach, there is a set of validation sample proportions, $P_1 = \frac{m_{\cdot 1}}{m}$, in which the resulting estimators from the 3S approach will be at least as efficient as the 2S approach. The 3S approach also gives investigators complete control over the categorical make-up of the validation sample, allowing sub-sampling costs in addition to efficiency to be considered when designing the validation sampling scheme. However, this set of P_1 values is unknown and is dependent on the observed information matrices.

In this section, we develop a numerical sample size determination algorithm that will assist in selecting the ideal value of P_1 (the value that minimizes the standard error of the MCB-corrected estimates under the 3S approach). This algorithm will approximate the relationship between the minimal value of m required to bound the half length of a Bonferonni interval for the parameter of interest with family wise confidence level 100 $(1-\alpha)\%$ and the range of possible P_1 values. The rationale for the use of Bonferroni intervals is that we may only wish to impose a width restriction on the intervals associated with the π-parameters exclusively, but may also seek to control of the overall error rate, necessitating the use of simultaneous confidence limit procedures.

As with most sample size determination algorithms, this algorithm will be developed using experimenter selected values for all probabilities ascertained via a pilot study, or possibly through an educated guess based on expert opinion. To handle the additional unknown quantities present when drawing a validation sample, we will use a simulation based approach to obtain their Monte Carlo estimates. Note that we can assume the original sample has already been drawn, so that the values of $n_{.1}$ and $n_{.0}$, $n = n_{.1} + n_{.0}$ are known. Finally, since the algorithm is motivated by outcome misclassification in electronic health record (EHR) data, we will assume for purposes of this discussion that the original sample size is sufficiently large to attain the requisite asymptotic properties, an assumption that is easily justified given the extremely large size of most EHR datasets.

In Section 3.2, we begin by reviewing the multi-sample asymptotic properties of the MLEs derived from the 2S and 3S approaches which are needed to derive simultaneous Bonferonni intervals. In Section 3.3, we develop the validation sample determination algorithm, along with the theoretical justification for its use. In

Section 3.4 we conduct a simulation study to demonstrate that the sample size determination algorithm functions as intended over a wide range of simulation parameters. We then generate sample size tables corresponding to the simulated examples in Chapter 2; we also present a number of additional sample size tables in Appendix D that have been validated though simulation as well to further generalize our sample size determination results.

3.2 Asymptotic Properties for Multi-Sample Maximum Likelihood Estimation

Recall the multi-sample framework characterizing the underlying probabilistic structure of the data as outlined in Section 2.3 and Appendix A. Under this framework, we demonstrated that the MLE, $\hat{\gamma}$, of the p-dimensional parameter vector, $\gamma = (\pi_{1|1}, \pi_{1|0}, \theta_{11}, \theta_{01}, \theta_{10}, \theta_{00})^T$, is asymptotically normally distributed with a variance-covariance matrix that can be written as the inverse of a sum of the probability weighted Fisher information matrices associated with each of the samples at the hypothesized value, γ_0,

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0) \xrightarrow{D} N\left(\mathbf{0}, \left[\sum_{s=1}^{s} \omega_s \mathcal{I}_s(\gamma_0)\right]^{-1}\right), \tag{3.1}$$

where $\hat{\gamma}_n \xrightarrow{p} \gamma_0$, $0 < \omega_s < 1$ and $\sum_s \omega_s = 1$. This asymptotic variance-covariance matrix can be estimated by,

$$\frac{1}{n}\sum_{s=1}^{S} n_s \bar{I}_s(\hat{\gamma}) \xrightarrow{p} \sum_{s=1}^{S} \omega_s \mathcal{I}_s(\gamma_0), \tag{3.2}$$

where $\left(\frac{n_1}{n}, ..., \frac{n_S}{n}\right) \longrightarrow (\omega_1, ..., \omega_s)$, $\bar{I}_s(\hat{\gamma}) = \frac{1}{n_s}\sum_{i=1}^{n_s} I_{si}(\hat{\gamma}) = \frac{1}{n_s}\sum_{i=1}^{n_s}\left[-\frac{\partial l_s(\gamma|x_{si})}{\partial\gamma\partial\gamma^T}\right]_{\gamma=\hat{\gamma}}$, $n = \sum_{s=1}^{S} n_s$, and n_s is the size of the s^{th} sample.

Thus, equation (3.2) allows us to write the following consistent estimators of the asymptotic variance-covariance matrices for the 2S approach,

$$\hat{\sigma}^2(\hat{\gamma}_2) = (n^u \bar{I}_O(\hat{\gamma}) + m\bar{I}_V(\hat{\gamma}))^{-1}, \tag{3.3}$$

and,

$$\hat{\sigma}^2(\hat{\gamma}_3) = [n\bar{I}_O(\hat{\gamma})+m_{.1}\bar{I}_{V_1}(\hat{\gamma})+m_{.0}\bar{I}_{V_0}(\hat{\gamma})]^{-1} = (n\bar{I}_O(\hat{\gamma})+m[P_1\bar{I}_{V_1}(\hat{\gamma})+(1-P_1)\bar{I}_{V_0}(\hat{\gamma})])^{-1}, \tag{3.4}$$

for the 3S approach, where the subscript O denotes the original sample, V denotes the validation sample in the 2S approach and $V_{\tilde{a}}$ denotes the validation sample associated with $\tilde{a} = 0, 1$ in the 3S approach. R code to compute each observed information matrix is provided in Appendix C.

3.3 Development of a Monte Carlo Sample Size Determination Algorithm

The purpose of this algorithm is to assist investigators in determining the minimal required validation sample size, m, needed to bound the half length of a Bonferroni interval with simultaneous confidence level $100(1-\alpha)\%$ by some value which we will denote as c. Further, we wish to consider the effect of altering the categorical makeup of the validation sample with respect to the outcome classification in the original sample, $\tilde{a} = 0, 1$. As noted in Section 3.2, we have a consistent estimator of the asymptotic variance-covariance matrix, equation (3.2), and asymptotic p-dimensional normality. Thus, we can create the approximate Bonferroni intervals with family-wise confidence level $100(1-\alpha)\%$ for γ,

$$b^T\hat{\gamma} \pm z_{\alpha/2p}\sqrt{b^T\left(\sum_{s=1}^{S} n_s\bar{I}_s(\hat{\gamma})\right)^{-1} b}, \tag{3.5}$$

where b is a vector of 1's and 0's that isolates the element in question, (e.g., if $b = (1, 0, 0, ...)$ then $b^T\gamma = \gamma_1$).

Recall that in this application, we are only interested in the parameters associated with the true underlying response probabilities and will treat the remaining parameters as nuisance parameters; as such, we will bound the largest half-interval associated with the parameters of interest by c. In other words, it is possible that interval lengths associated with the nuisance parameters will surpass the bounds on the response probabilities for the chosen value of m; however, we will not consider bounding these. Note that there are two parameters of interest in this application, $\pi_{1|1}$ and $\pi_{1|0}$; we will choose to bound the larger of the two simultaneous intervals as this will also bound the shorter.

For the 2S approach we have,

$$z_{\alpha/2p}\sqrt{b^T\left(n^u\bar{I}_O(\hat{\gamma}) + m\bar{I}_V(\hat{\gamma})\right)^{-1}b} \leq c, \tag{3.6}$$

and

$$z_{\alpha/2p}\sqrt{b^T\left(n\bar{I}_O(\hat{\gamma}) + m^*[P_1\bar{I}_{V_1}(\hat{\gamma}) + (1-P_1)\bar{I}_{V_0}(\hat{\gamma})]\right)^{-1}b} \leq c, \tag{3.7}$$

for the 3S approach using the second expression in equation (3.4). To differentiate between the validation sample size for the 2S and 3S approaches, we will refer to the 3S approach with an asterisk (*) for this section.

Using equations (3.6) and (3.7), we can solve for m and m^*, for chosen values of c and P_1, using a root finder such as *uni.root*() in R. However, we must first have an idea as to how the negative of the Hessian matrices will behave for each subsample. As mentioned in Section 3.1, this algorithm will be developed assuming specification of the true parameter values based on some form of prior knowledge such as a pilot study or educated guess. Since the MLEs are unbiased estimators of these true values, we can evaluate the Hessians at the specified (experimenter chosen) parameters. Therefore, we are able to calculate a value for the matrix, $\bar{I}_O(\gamma)$, since the $n_{d\tilde{a}}$'s are known. However, more information is needed to approximate the other matrices.

Each of the observed information components characterizing the validation samples, $\bar{I}_V(\gamma)$, $\bar{I}_{V_1}(\gamma)$ and $\bar{I}_{V_0}(\gamma)$, are a function of the realized values of $m_{d\tilde{a}}$ and $y_{d\tilde{a}}$. In other words, these quantities are subject to sampling variability and as such we will apply a Monte Carlo based method of estimation. For the k^{th} Monte Carlo iteration, we can utilize the investigator-specified parameters and generate a dataset with binary outcome misclassification. We then draw a validation sample of size m using the sampling method associated with the 2S and 3S approaches and observe the values of $m_{d\tilde{a}}^{(k)}$ and $y_{d\tilde{a}}^{(k)}$. Using these values, we can calculate the remaining observed information matrices, $\bar{I}_V^{(k)}(\gamma)$, $\bar{I}_{V_1}^{(k)}(\gamma)$ and $\bar{I}_{V_0}^{(k)}(\gamma)$. Repeating this procedure a large number of times, $B = 1,000$, will give us their MC estimates, eg. $\bar{I}_V^{MC}(\gamma) = \frac{\sum_{k=1}^{B} \bar{I}_V^{(k)}(\gamma)}{B}$. These values can then be used in equations (3.6) and (3.7) allowing us to solve for m and m^*.

The final problem is in the selection of m, $m_{.1}$ and $m_{.0}$ in the simulation step. The size of the validation sample used to derive the MC estimates will effect the resulting $\bar{I}(\gamma)$'s. However, these matrices are consistent estimates of the s-sample Fisher information matrices as noted in Theorem A.2.4 Appendix A,

$$\bar{I}_s(\gamma) = \frac{1}{n_s} \sum_{i=1}^{n_s} \left[-\frac{\partial l_s(\gamma|x_{si})}{\partial\gamma\partial\gamma^T} \right] \xrightarrow{p} E[-\frac{\partial^2 l_s(\gamma)}{\partial\gamma\partial\gamma^T}] = \mathcal{I}_s(\gamma), \tag{3.8}$$

for all samples, $s = 1, ..., S$. Thus, for the s^{th} sample, the approximation will improve as the subsample size n_s increases, as we have a weak law of large numbers for each sample's observed information matrix. Assessing the convergence behaviour of these matrices as m increases, $m < n$, will allow us to select a reasonable validation sample size from which to generate in the Monte Carlo step of the algorithm.

To simplify the discussion, let P_m denote the proportion of observations in the original sample that will be selected as the validation sample, $m = P_m n$. We can use numerical methods to observe the behaviour of the $\bar{I}_s(\gamma)$'s as $m \longrightarrow n$. We will use the p^{th} ($p = 6$) root of the determinant of the matrix to assess the convergence behaviour of the negative of the Hessians as the subsample sizes increase. In design of experiments, this is used as a criterion for optimality of design (see Montgomery (2005) [31]). Since we are assessing the change in information relative to an increasing value of m, stability of this measure should be adequate to establish convergence of these matrices.

For each iteration of the 1000 runs of each simulation study, for chosen values of n, m and γ, we will generate an EHR data set and draw validation samples under the 2S and 3S approaches. Using these observed values, we will generate 1000 values of the observed information components and derive the Monte Carlo estimates for each. Next, we will calculate the 6^{th} root of the determinant of these matrices and repeat this procedure for fixed n and γ over increasing values of m. We will show that these matrices appear to attain their limit for values of P_m that are relatively low.

For the observed information matrices associated with the 3S approach, we must also have some idea of the value of P_1 since for every m, the values of $m_{.1}$ and $m_{.0} = m - m_{.1}$ may affect the limiting behaviour of the s-sample observed information matrices. A reasonable starting point is $P_1 = \frac{n_{.1}}{n}$ as this would generate validation samples of similar categorical make-ups for both approaches. However, we will also alter P_1 to observe its impact on convergence. We will generate samples from 2 additional values of P_1, namely $P_1 = 0.5 \times \frac{n_{.1}}{n}$ and $P_1 = 1.5 \times \frac{n_{.1}}{n}$ (note that we specify target simulation parameters that produce observed incidence rates that will keep $P_1 < 1$). To compare these two sets of results, we will take the differences between the resulting MC estimates of the information matrices generated via each value of P_1. Plotting these with respect to an increasing P_m will further allow us to visualize the relative difference in the limiting behaviour for different values of P_1.

We will use the same sets of simulation parameters listed in Table 2.3.1 of Chapter 2 for presentation of the sample size results in this discussion; however, many other sets of parameters were run with similar results as can be seen in Appendix D. These sets of parameters were designed to produce increasing incidence rates of category 1 observations, $\frac{n_{.1}}{n}$, in the original sample, O. This will allow us to further observe the effect on convergence associated with an increasing incidence rate. Results are presented in Table 3.3.1. Note that the column marked '$E(n_{.1})$' denotes the expected count of category 1 observations, which is described in Chapter 2 as $n_{1.}P(\tilde{A} = 1|D = 1) + n_{0.}P(\tilde{A} = 1|D = 0) = n_{1.}(\pi_{1|1}\theta_{11} + (1 - \pi_{1|1})\theta_{10}) + n_{0.}(\pi_{1|0}\theta_{01} + (1 - \pi_{1|0})\theta_{00})$ where $n_{d.} = n_{d1} + n_{d0}$ denotes the exposure (drug) group sample size in the original data. We will select $n = 10,000$ and $n = 100,000$ and will slowly increase P_m across the interval $(0, 0.5]$. Finally, the exposure group sizes were held equal in the figures presented below at $\frac{n}{2}$ in each exposure category, $d = 0, 1$. Additional simulations were run to account for differing exposure group sizes and produced similar results; hence they are not presented here.

Table 3.3.1: Chosen target parameters for the simulation studies comparing the 2S and 3S approaches.

ID	$E(n_{.1})$	$\pi_{1\|1}$	$\pi_{1\|0}$	θ_{11}	θ_{01}	θ_{10}	θ_{00}
1	$0.097n$	0.05	0.10	0.85	0.95	0.05	0.01
2	$0.195n$	0.1	0.20	0.95	0.85	0.05	0.10
3	$0.300n$	0.3	0.20	0.90	0.90	0.10	0.10

Figure 3.3.1: Convergence behaviour of the validation subgroup's observed information for parameter set 1.

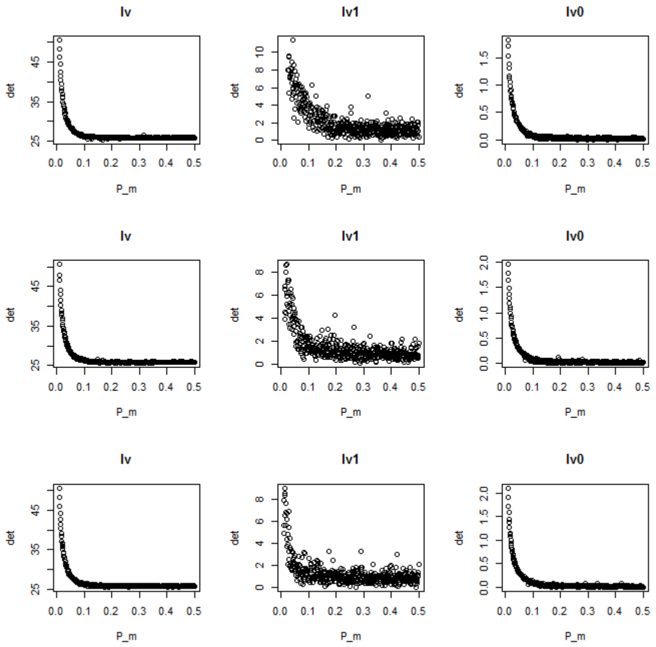


Figure 3.3.2: Convergence behaviour of the validation subgroup's observed information for parameter set 2.

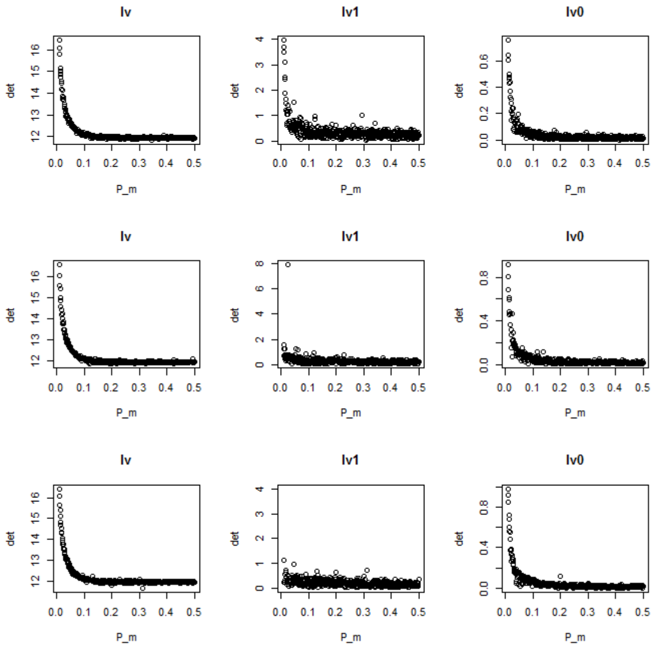
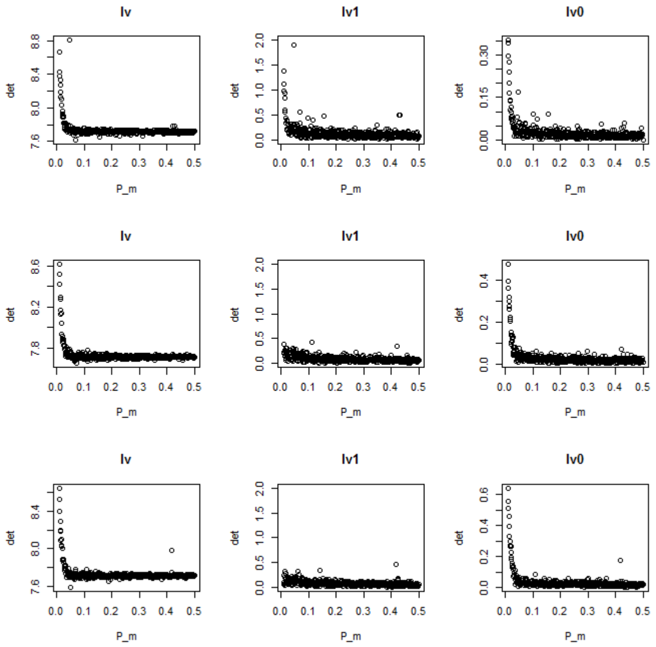


Figure 3.3.3: Convergence behaviour of the validation subgroup's observed information for parameter set 3.



It seems clear from the results presented in Figures 3.3.1 - 3.3.3, that a fairly low value of P_m is needed for the matrices to converge. All these figures are based on the $n = 10,000$ simulations; as the $n = 100,000$ simulations produced virtually identical results and are omitted. Across all 3 sets of simulation parameters the plots demonstrate that the determinants of these matrices seem to stabilize at approximately $P_m = 0.2$ or 0.3. For some of the $\bar{I}_{V_1}(\gamma)$ plots, the proximity of the MC estimates across all values of P_m is extremely close. Thus, the range of the y-axis was altered to provide plots that are comparable graphically, since selecting the minimum and maximum of the points (R's default) produced what appeared to be a random spread of points.

Next, all the figures demonstrate similar behaviour, which implies that the incidence rates within the underlying sample have little effect on convergence. Finally, upon first glance, it appears that generation from different values of P_1 has no effect as well. Note, that rows of plots in each figure are the results associated with $P_1 = 0.5 \times \frac{n_{.1}}{n}$, $P_1 = \frac{n_{.1}}{n}$ and $P_1 = 1.5 \times \frac{n_{.1}}{n}$, in that order. To verify this, we will plot their differences across the values of P_m in the plots in

Figures 3.3.4-3.3.6. The first row is the difference between $P_1 = 0.5 \times \frac{n_{.1}}{n}$ and $P_1 = \frac{n_{.1}}{n}$, the second row is between $P_1 = \frac{n_{.1}}{n}$ and $P_1 = 1.5 \times \frac{n_{.1}}{n}$ and the third is between $P_1 = 1.5 \times \frac{n_{.1}}{n}$ and $P_1 = 0.5 \times \frac{n_{.1}}{n}$.

Figure 3.3.4: Convergence behaviour of observed information due to validation sampling differences for parameter set 1.

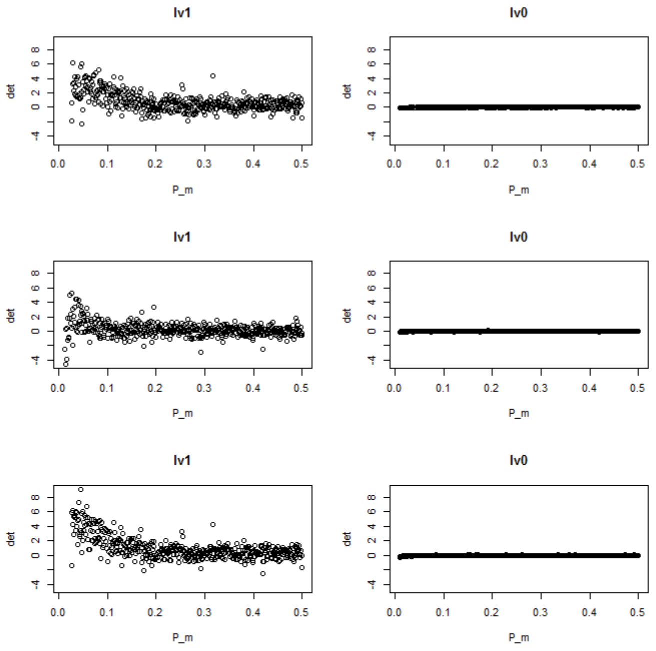


Figure 3.3.5: Convergence behaviour of observed information due to validation sampling differences for parameter set 2.

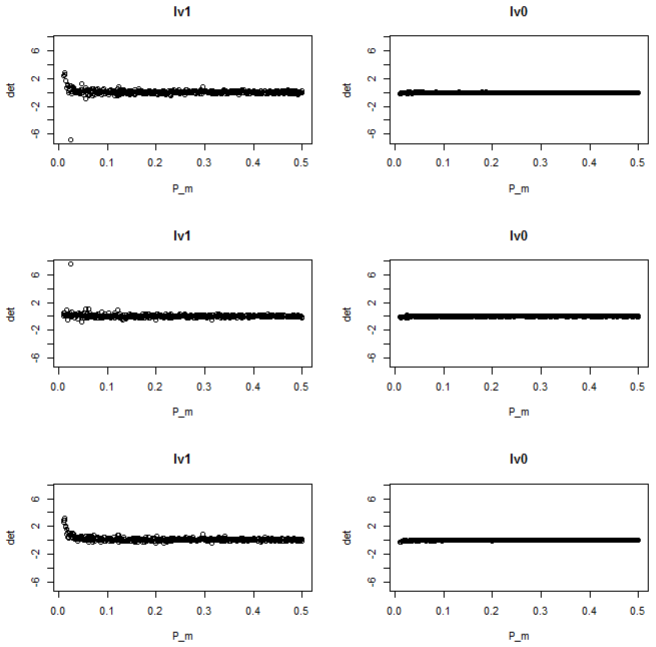
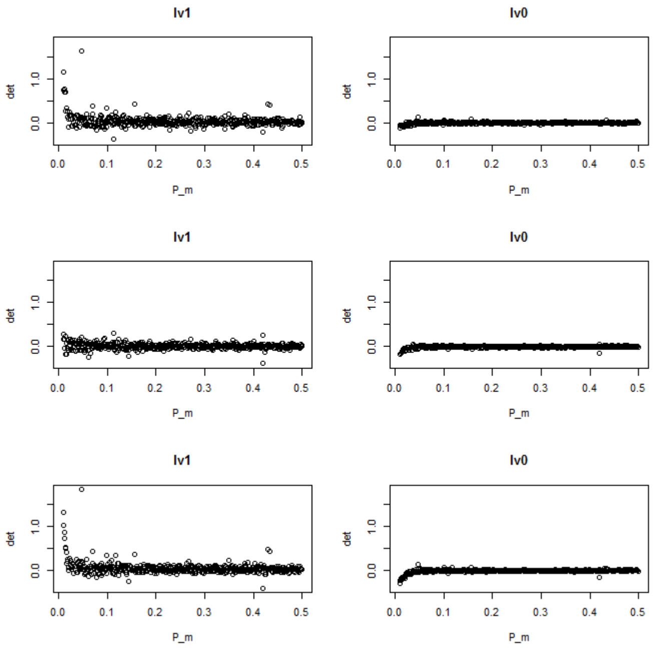


Figure 3.3.6: Convergence behaviour of observed information due to validation sampling differences for parameter set 3.



These plots demonstrate that across all of the differences, convergence to 0 occurs at reasonable values of P_m. Note that we fixed the y-axis range based on the minimum and maximum across all differences for each parameter set.

These results are as expected since the validation sample information matrices are a function of parameters that are specified by the investigator. Thus, the only convergence issues are based on the realized sample counts, which are drawn from binomial and multinomial random variables, as pointed out in Section 2.3. Hence, the values of $m_{d\tilde{a}}$ and $y_{d\tilde{a}}$ will tend towards their expected counts over many iterations, and the influence of m and P_1 will be to provide the 'number of trials' parameter for each possible subgroup in the validation sample. The categorization probability parameters are assumed known which implies that as long as $m, m_{\cdot 1}$ and $m_{\cdot 0}$ are large enough to enable estimation of the expected counts in each category, we should see Monte Carlo estimates that tend towards the validation sample Fisher information matrices.

Returning to the Monte Carlo step of the validation sample size determination algorithm, we can safely recommend using a $P_m = 0.3$ or 0.4, with the choice of P_1 appearing to be of little importance. We will select $P_1 = \frac{n_{.1}}{n}$ for sampling convenience, in that we will run less risk of attempting to draw a category 1 validation sample size that exceeds the possible number of observations with that category in the original sample. Finally, we have specified a manner of generating Monte Carlo estimates of the observed information components at the true parameter values for both 2S and 3S approaches that will reasonably approximate the s-sample Fisher information components, $s = 1, ..., S$. Hence, we are able to summarize the validation sample size determination algorithm in 2 steps.

Algorithm 1: Generation of $\bar{I}_O$, $\bar{I}_V(\gamma)$, $\bar{I}_{V_1}(\gamma)$, $\bar{I}_{V_0}(\gamma)$

```
i) Select $n, \gamma, P_m, P_1$ where $P_m = \frac{m}{n} \in [0.3, 0.5]$ and $P_1 = \frac{m_{.1}}{m}$ with a recommendation of $P_1 = \frac{n_{.1}}{n}$.

ii) FOR LOOP; indexed by k

        - Generate a dataset with binary outcome misclassification (see Appendix C for R code or Chapter 2)

        - Calculate $\bar{I}_O^{(k)}(\gamma)$

        - Draw a validation sample of size $m = P_m \times n$ from generated dataset

        - Calculate $\bar{I}_V^{(ki)}(\gamma)$

        - Draw a validation sample from the category 1 observations of size, $P_1 \times m$

        - Calculate $\bar{I}_{V_1}^{(ki)}(\gamma)$

        - Draw a validation sample from the category 2 observations of size, $(1 - P_1) \times m$

        - Calculate $\bar{I}_{V_0}^{(k)}(\gamma)$

    END LOOP
```

$\bar{I}_O^{(MC)}(\gamma)$=$\sum_k \bar{I}_O^{(k)}(\gamma)$
$\bar{I}_V^{(MC)}(\gamma)$=$\sum_k \bar{I}_V^{(k)}(\gamma)$
$\bar{I}_{V_1}^{(MC)}(\gamma)$=$\sum_k \bar{I}_{V_1}^{(k)}(\gamma)$
$\bar{I}_{V_0}^{(MC)}(\gamma)$=$\sum_k \bar{I}_{V_0}^{(k)}(\gamma)$
This algorithm will produce Monte Carlo estimates for each set of n, γ, P_m, P_1.

Algorithm 2: 2S Approach Sample Size Determination Algorithm

i) For $n, \gamma, \bar{I}_O^{(MC)}(\gamma), \bar{I}_C^{(MC)}(\gamma)$

ii) Select c.

iii) Numerically solve $z_{\alpha/2p}\sqrt{b^T\left(n^u\bar{I}_O(\hat{\gamma}) + m\bar{I}_V(\hat{\gamma})\right)^{-1}b} = c$ for $m^{(1)}$ where $b = (1,0,0,0,0,0)$ followed by $m^{(2)}$ where $b = (0,1,0,0,0,0)$.

iv) $m = max(m^{(1)}, m^{(2)})$.

Algorithm 3: 3S Approach Sample Size Determination Algorithm

i) Select $n, \gamma, \bar{I}_O^{(MC)}(\gamma), \bar{I}_{V_1}^{(MC)}(\gamma), \bar{I}_{V_0}^{(MC)}(\gamma)$

ii) Select c and P_1.

iii) Numerically solve $z_{\alpha/2p}\sqrt{b^T\left(n\bar{I}_O(\hat{\gamma}) + m(P_1\bar{I}_{V_1}(\hat{\gamma}) + (1-P_1)\bar{I}_{V_2}(\hat{\gamma})\right)^{-1}b} = c$ for $m^{(1)}$ where $b = (1,0,0,0,0,0)$ followed by $m^{(2)}$ where $b = (0,1,0,0,0,0)$.

iv) $m^* = max(m^{(1)}, m^{(2)})$.

In the following section, we will produce validation sample size determination tables corresponding with the sets of parameters in Table 3.3.1. In Section 2.3.1 simulation studies were conducted on these parameter sets which confirm that the algorithms are working well.

3.4 Simulation Study to Investigate Monte Carlo Sample Size Determination Algorithm

In Tables 2.3.2a, 2.3.3a and 2.3.4a in Section 2.3.1, the standard error estimates were presented for the three sets of parameters. We reproduce these here for convenience.

Table 3.4.1: Simulation study results for the 2S versus 3S approach for parameter set 1.

(a) Monte Carlo Estimates of Standard Error.

		3S: P_1							
	2S	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
$\hat{\sigma}(\hat{\pi}_{1\|1})$	0.00795	0.00787	0.00655	0.00616	0.00622	0.0065	0.00722	0.0085	0.01129
$\hat{\sigma}(\hat{\pi}_{1\|0})$	0.00643	0.00649	0.00597	0.00597	0.00617	0.00649	0.00718	0.0084	0.01107
max	0.00795	0.00787	0.00655	0.00616	0.00622	0.0065	0.00722	0.0085	0.01129

Table 3.4.2: Simulation study results for the 2S versus 3S approach for parameter set 2.

(a) Monte Carlo Estimates of Standard Error.

		3S: P_1							
	2S	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
$\hat{\sigma}(\hat{\pi}_{1\|1})$	0.00904	0.01178	0.00906	0.00793	0.0075	0.0074	0.00768	0.00849	0.01046
$\hat{\sigma}(\hat{\pi}_{1\|0})$	0.01342	0.01659	0.01334	0.0124	0.01209	0.01233	0.01299	0.01408	0.01682
max	0.01342	0.01659	0.01334	0.0124	0.01209	0.01233	0.01299	0.01408	0.01682

Table 3.4.3: Simulation study results for the 2S versus 3S approach for parameter set 3.

(a) Monte Carlo Estimates of Standard Error.

		3S: P_1							
	2S	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
$\hat{\sigma}(\hat{\pi}_{1\|1})$	0.01376	0.01986	0.01543	0.01393	0.01309	0.0128	0.01341	0.01382	0.01635
$\hat{\sigma}(\hat{\pi}_{1\|0})$	0.01275	0.01923	0.01487	0.01285	0.01215	0.01166	0.01174	0.01199	0.01397
max	0.01376	0.01986	0.01543	0.01393	0.01309	0.0128	0.01341	0.01382	0.01635

These estimates are the results of simulation studies with $n = 10,000$ and $m = m^* = 1,000$, in which for each iteration, a dataset with binary outcome misclassification was generated (as in Appendix C) and the 2S and 3S approaches were applied to produce MCB-adjusted estimates for the π-parameters. The inverses of the observed information matrices were calculated to obtain the standard error estimates. Note, that under the 3S approach, for each dataset, 8 validation samples were drawn, one for each value of $P_1 \in (0.1, ..., 0.8)$. Finally, the row marked 'max' denotes $\max(\hat{\sigma}(\hat{\pi}_{1|1}), \hat{\sigma}(\hat{\pi}_{1|0}))$. Each study was based on 1,000 iterations.

Utilizing the validation sample size determination algorithms presented earlier, we generate the following sample size tables. Note that $P_1 = 0.9$ was excluded from the previous simulation work due to possible sub-sample size issues, however, we include the Monte Carlo sample size determination results here.

Table 3.4.4: Monte Carlo sample size determination for parameter set 1 with n= 10,000.

P_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	2S
c	m^*	m^*	m^*	m^*	m^*	m^*	m^*	m^*	m^*	m
0.01	-	-	-	-	-	-	-	-	-	5498
0.02	1074	707	617	609	650	746	931	-	-	1107
0.03	461	304	265	261	279	320	400	568	-	475
0.04	256	169	148	145	155	178	222	316	605	264
0.05	163	108	94	93	99	114	142	201	385	168
0.06	113	75	65	64	69	79	98	140	267	117
0.07	83	55	48	47	51	58	72	102	196	86
0.08	64	42	37	36	39	44	55	78	150	66
0.09	50	33	29	29	31	35	44	62	118	52
0.1	41	27	24	23	25	29	36	50	96	42
0.11	34	22	20	19	21	24	29	42	79	35
0.12	29	19	17	16	17	20	25	35	67	29
0.13	24	16	14	14	15	17	21	30	57	25
0.14	21	14	12	12	13	15	18	26	49	22
0.15	18	12	11	11	11	13	16	23	43	19

Table 3.4.5: Monte Carlo sample size determination for parameter set 2 with n= 10,000.

P_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	2S
c	m^*	m^*	m^*	m^*	m^*	m^*	m^*	m^*	m^*	m
0.01	-	-	-	-	-	-	-	-	-	-
0.02	6269	3897	3252	3093	3206	-	-	-	-	3952
0.03	2319	1442	1203	1144	1186	1327	1621	2265	-	1460
0.04	1232	766	640	608	631	705	862	1204	-	775
0.05	769	478	399	380	394	440	538	751	1416	484
0.06	527	328	274	260	270	302	369	515	970	332
0.07	384	239	200	190	197	220	269	375	707	242
0.08	293	182	152	145	150	168	205	286	539	184
0.09	231	144	120	114	118	132	161	225	424	145
0.1	186	116	97	92	96	107	130	182	343	117
0.11	154	96	80	76	79	88	108	150	283	97
0.12	129	81	67	64	66	74	90	126	238	81
0.13	110	69	57	55	57	63	77	107	202	69
0.14	95	59	49	47	49	55	66	93	174	60
0.15	83	52	43	41	43	48	58	81	152	52

Table 3.4.6: Monte Carlo sample size determination for parameter set 3 with n= 10,000.

P_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	2S
c	m^*	m^*	m^*	m^*	m^*	m^*	m^*	m^*	m^*	m
0.01	-	-	-	-	-	-	-	-	-	-
0.02	-	6716	5217	4652	4549	4826	-	-	-	5209
0.03	3630	2082	1618	1442	1411	1496	1741	2325	-	1616
0.04	1847	1060	823	734	718	761	886	1183	2139	822
0.05	1132	650	505	450	440	467	543	725	1311	504
0.06	768	441	343	306	299	317	369	492	890	342
0.07	557	320	249	222	217	230	267	357	645	248
0.08	423	243	189	168	165	175	203	271	490	189
0.09	332	191	148	132	129	137	160	213	385	148
0.1	268	154	120	107	105	111	129	172	311	120
0.11	221	127	99	88	86	91	106	142	256	99
0.12	185	107	83	74	72	77	89	119	215	83
0.13	158	91	71	63	62	65	76	101	183	70
0.14	136	78	61	54	53	56	65	87	157	61
0.15	118	68	53	47	46	49	57	76	137	53

Upon first look, the convex relationship discussed earlier between the standard error and P_1 appears evident. It also seems as though there is clearly a range of values of P_1 in which the 3S approach is more desirable. To compare these tables with the simulations from Section 2.3.1 we will select values of m close to 1,000 and observe the value of c bounding the half interval. Then, we can solve the actual half interval length using the estimates in Tables 3.4.1, 3.4.2 and 3.4.3 and computing $z_{\alpha/(2\times 6)} \times \max(\hat{\sigma}(\hat{\pi}_{1|1}), \hat{\sigma}(\hat{\pi}_{1|0}))$. For $\alpha = 0.05$, the z-score will be $z_{0.05/12} = 2.638$.

For parameter set 1, the values of m^* in Table 3.4.4 decrease quickly. However, for P_1 just above 0.1, we see a value of $m^* \approx 1000$. For the 2S approach, the closest we come is $m = 1107$. Both entries are associated with a value of $c = 0.02$. Thus, we would expect to see both the 2S and 3S intervals be slightly longer than 0.02. Referring to Table 3.4.1, the 2S estimate of standard error is 0.00795, giving us a half length of 2.638 × 0.00795 = 0.02097; for the 3S approach, at $P_1 = 0.1$ we have an estimate of 0.00787, giving us a half length of 2.638 × 0.00787 = 0.02076.

To give one more example, consider Table 3.4.5. For the 3S approach at $P_1 = 0.4$, we would expect the half interval to be just over c =0.03 and for the 2S approach, somewhere between c =0.03 and 0.04. Referring to Table 3.4.2, at $P_1 = 0.4$ the half length is 2.638 × 0.01209 = 0.0319 for the 3S approach and for the 2S approach it is 2.638 × 0.01342 = 0.0354. Thus, both examples produce results that would be expected, further validating the sample size determination algorithms.

Finally, to address the range of values of P_1 in which the 3S approach outperforms the 2S approach, we will plot the resulting values of m and m^* against P_1 for each of the sets of parameters in Table 3.3.1. Since these plots will change with respect to c, we will select $c = 0.05$ and 0.1 in order to observe the resulting change in behaviour.

Figure 3.4.1: Plot of validation sample size relative to 2S and 3S sampling approaches for parameter set 1. Two bounds are considered, $c = 0.05, 0.1$ with an expected incidence rate of 0.097.

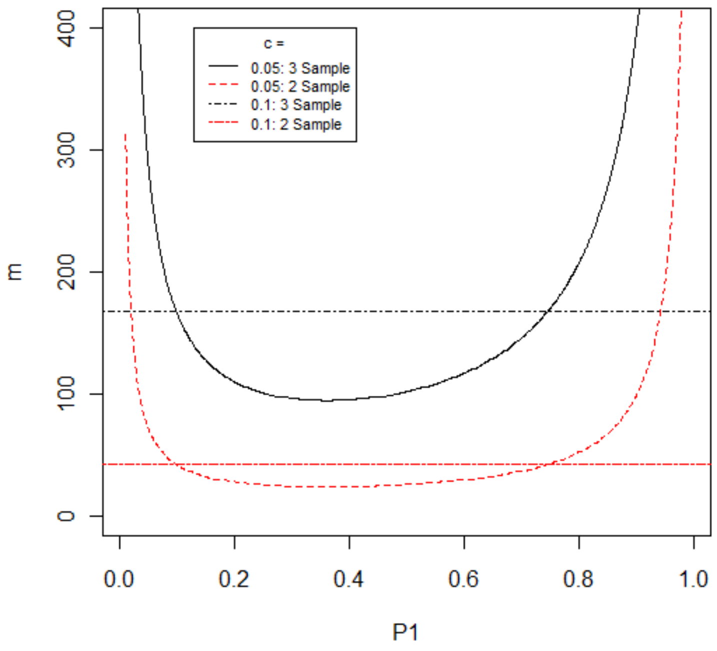


Figure 3.4.2: Plot of validation sample size relative to 2S and 3S sampling approaches for parameter set 2. Two bounds are considered, $c = 0.05, 0.1$ with an expected incidence rate of 0.195.

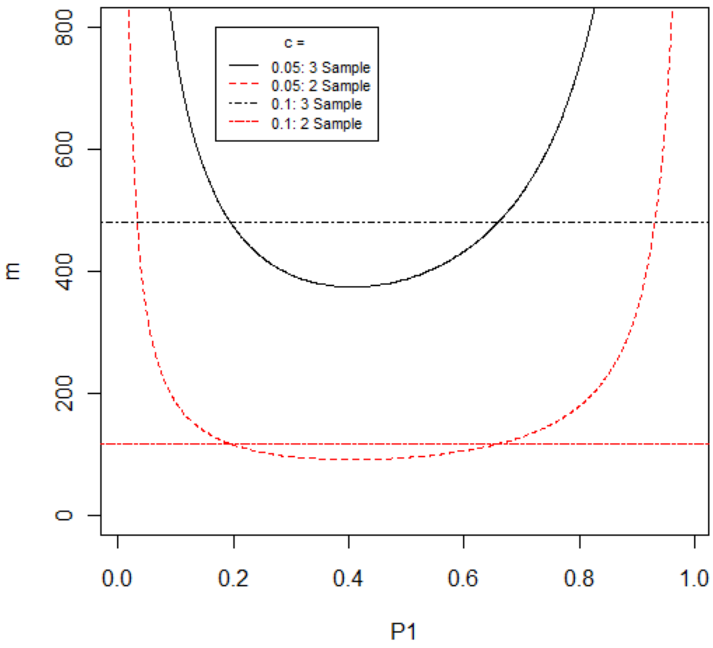
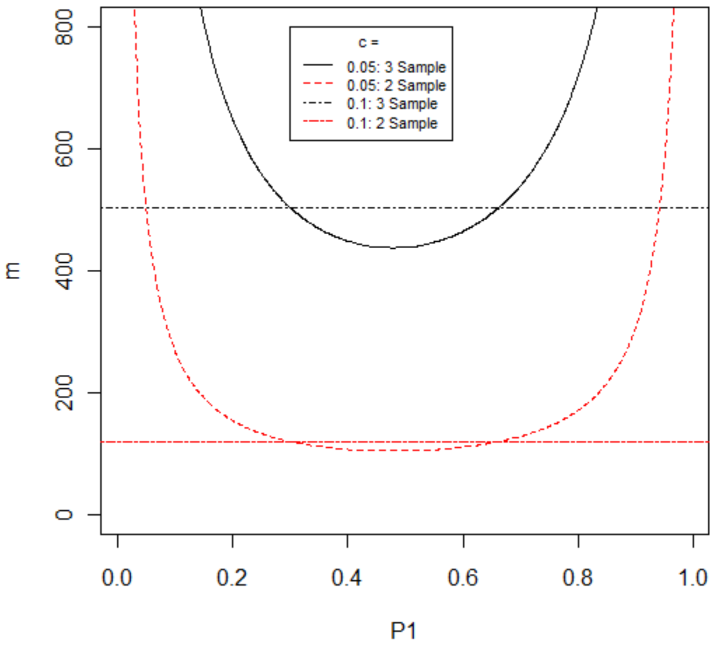


Figure 3.4.3: Plot of validation sample size relative to 2S and 3S sampling approaches for parameter set 3. Two bounds are considered, $c = 0.05, 0.1$ with an expected incidence rate of 0.3.



These plots confirm that there is a range of P_1 values in which the 3S approach is ideal with respect to the minimal validation sample size needed to attain a bound of $c = 0.05$ and 0.1. Also, it appears that the lower intersection point between the 2S and 3S approaches occurs close to the observed incidence rate. Finally, the implication of these plots is that we are able to reasonably recommend which approach to use, with the approximate value of P_1 that requires the smallest m^* in the 3S approach, as well as all the requisite information to assess the trade-off between validation sample size and variance. For example, in parameter set 2 (Figure 3.4.2), we see that the minimum occurs at approximately $P_1 = 0.4$, which, according to Table 3.4.5, requires a value of $m^* = 380$ for a bound of $c = 0.05$ and $m^* = 92$ for a bound of $c = 0.1$. The associated values of m are $m = 484$ for $c = 0.05$ and $m = 117$ for $c = 0.1$. Thus, the ideal approach to minimize the required validation sample size is to select a sample of $m_{.1} = 380 \times 0.4 = 152$ and $m_{.0} = 228$ for the bound $c = 0.05$ and $m_{.1} = 92 \times 0.4 \approx 37$ and $m_{.0} = 55$ for $c = 0.1$.

These results also have implications with respect to differing costs of drawing validation samples in each category. For instance, there may be an AE for which a simple lab test can confirm an erroneous diagnosis, although an expensive test must be used to validate that the diagnosis was accurate. In such cases, we would ideally wish to minimize the number of observations required in the validation sample in category 1, $m_{.1}$. These plots help to assess this trade-off: for instance for parameter set 3 (Figure 3.4.3), the 2S and 3S approaches appear to require the same validation sample size at approximately $P_1 = 0.3$. This is the expected incidence rate, and as such, the 2S approach would likely select approximately 30% of the validation sample within category 1. Referring to Table 3.4.6, we need $m^* = 505$ and $m = 504$ observations at this value to attain a bound of $c = 0.05$, which implies that we are sampling approximately (exact for the 3S approach) 151 observations from category 1 in our validation samples. However, if we are willing to sacrifice some precision and increase the minimal required m^*, we could select $P_1 = 0.2$. This would require $m^* = 650$ and $m_{.1} \approx 130$, and save the cost of 21 category 1 observations, while adding the cost of sampling 166 observations from category 2.

Note that the crux of this algorithm is based on numerically generated datasets and, as such, the resulting entries in Tables 3.4.4, 3.4.5 and 3.4.6 will vary slightly due to simulation error. The simplicity of generating binary data with the presence of outcome misclassification is crucial for the justification of the use of this algorithm. The R code used in these sample size calculations is presented in Appendix C. Additional simulations are shown in Appendix D producing results similar to those presented in this chapter.

3.5 Conclusions and Discussion

The present sample size determination algorithm was developed for the purpose of providing practitioners a practical method of selecting between the 2S and 3S approaches, and to provide guidance on how to select the validation sample. This algorithm approximates the relationship between the validation sample size, m, the categorical make-up of the validation sample in the 3S approach, P_1, and the width of a Bonferroni interval, c, for the parameters of interest. We validated this approximation through Monte Carlo simulation under a wide range of conditions, including those presented in Appendix D.

The key assumption underlying this work is that binary outcome misclassification is adequately described by the data generation algorithm, given that the experimenter provides reasonable prior information on the values of the true response probabilities. The data generation portion of the algorithm is essential in that we must account for the realized subgroup counts in each of the validation samples in both the 2S and 3S approaches. However, since we are altering the value of a binary classifier at a rate that is assumed to be known, we can be confident that the resulting Monte Carlo estimates will be representative. In other words, the simplicity of conducting a Bernoulli trial with a known parameter over a large number of iterations is sufficient justification to be confident about the data generation portion of this algorithm.

Nonetheless, these results depend on the prior values selected for the underlying response probabilities; if these values are inaccurate, the sample size determination results may be misleading. Future research on the impact of misspecification on the approximation of the relationship between m, P_1 and c is therefore warranted. We may be satisfied with achieving a validation sampling scheme that, although not optimal in terms of the efficiency of estimation relative to P_1, represents an improvement over the 2S approach; this may well be the case if we are in a situation in which the cost of drawing the validation sample observ0ations from each category differs markedly.

Chapter 4

On the Optimization of a Validation Sampling Approach for Misclassification Bias Adjustment in Logistic Regression Models

4.1 Introduction

In this chapter we extend our previous work in misclassification bias (MCB) adjusted estimation, to include additional covariates included within electronic health record (EHR) data. As discussed in Chapter 1, EHRs are used to record patient data associated with the care they received within their health care institutions. As such, they contain socio-demographic patient information as well as the information gathered during such an encounter. Electronic health records can be used for "setting objectives and planning patient care, documenting the delivery of care and assessing the outcome of care" (Häyrinen et al. (2008) [16]). Thus, the use of these records to address drug safety concerns is limited to information that is recorded during an encounter or generated by an encounter with the health care system.

That said, there exists a large amount of additional information beyond AE and drug utilization status included with every entry in these data sets. This information can be used to investigate the impact of these variables on the probability of adverse event (AE) occurrence. Thus, we will extend the work in Chapter 2 to incorporate this additional information by modelling the π and θ-parameters via logistic regression. Using the previously introduced multi-sample framework we can then derive MCB-adjusted maximum likelihood estimators (MLEs) under both the 2S (random sampling) and 3S (conditional sampling) approaches. Recall that Lyles et al. (2011) [25] present a likelihood that is a form of the general likelihood presented in Carroll et al. (2006) [4]. This likelihood mirrors the 2S approach of Chapter 2 in that the validation sample is drawn randomly. As such we will present this likelihood using our notation as well as introduce the 3S likelihood.

Recall that under the 2S approach the validation sample will probabilistically mirror the original EHR data with respect to its categorical composition. However, the resulting validation data needs to contain adequate information on four outcome categories, namely, $\{\tilde{A} = \tilde{a}, A = a\}$, $\tilde{a}, a = 0, 1$. The added complexities associated with the additional covariate information hypothetically influencing the misclassification rates, leads to added difficulties in drawing adequate validation sample sizes in each of these subgroups. While the 3S approach allows some control over the $\tilde{a}$ subgroup sample sizes, extreme scenarios regarding the influence of covariates on misclassification within these categories may nevertheless present difficulties.

We will begin by presenting the 2S approach and extending the 3S approach from Chapter 2 to incorporate additional variables of interest. We then outline the multi-sample framework and present the asymptotic properties of the bias-adjusted estimators under both approaches. The finite sample properties of the adjusted estimators will be investigated through Monte Carlo simulation. We will investigate the estimated asymptotic relative efficiency (ARE) of the 2S and 3S approaches for differing values of $P_1 = \frac{m_1}{m}$. We will purposefully select simulation target parameters that will push the sample size boundaries within limits reflective of realistic EHR datasets. Since it seems reasonable that outcome misclassification in EHR data may depend on a variety of patient characteristics comprised of differing data types, we will model the θ-parameters using both discrete and continuous covariates in these simulation studies. Recall that Lyles et al. (2011) [25] considered the use of internal validation data with differential misclassification based solely on the 2S approach and considered target parameters designed to mimic a specific data example prompting the desire for more extensive simulations over a wide range of target parameters. Next, we will demonstrate that selection of either the 2S or the 3S approach is more complicated in the presence of this additional explanatory information. In fact, it will be seen that the conclusion presented in Chapter 2, that the 3S approach will be at least as good an estimator as the 2S approach provided that the ideal value of P_1 is used, will not necessarily hold. Hence, the need for a validation sample size determination algorithm is clear. We conclude this discussion by extending the sample size determination algorithms of Chapter 3 to the present context, and discuss some of the potential difficulties in their application and generalization.

4.2 Binomial Likelihood with Outcome Misclassification and Logistic Regression

In this section we will extend the binomial likelihood in the presence of outcome misclassification to account for the covariate information. To do so, we can write, $\pi_d(z_i) = P(A = 1|D = d, Z = z_i)$ and $\theta_{da}(z_i) = P(\tilde{A} = 1|A = a, D = d, Z = z_i)$ where z_i denotes the covariate vector for the i^{th} patient. We will model the data under a multi-sample framework, using the combined likelihood contributions of the original EHR data with the internally gathered validation data in the manner discussed in Section 2.2. The components of the original EHR data likelihood,

L_O, can be written as,

$$\begin{aligned} P[\tilde{A}=1|A=a,D=d,Z=z_i] &= \pi_d(z_i)\theta_{d1}(z_i)+(1-\pi_d(z_i))\theta_{d0}(z_i) \qquad (4.1) \\ P[\tilde{A}=0|A=a,D=d,Z=z_i] &= 1-\pi_d(z_i)\theta_{d1}(z_i)-(1-\pi_d(z_i))\theta_{d0}(z_i), \end{aligned}$$

for $i=1,...,n$, $d,a=0,1$.

Thus, the observed EHR data portion of the likelihood is,

$$L_O=\prod_{d=0,1}\prod_{i=1}^{n}[\pi_d(z_i)\theta_{d1}(z_i)+(1-\pi_d(z_i))\theta_{d0}(z_i)]^{\tilde{a}_{di}}[1-\pi_d(z_i)\theta_{d1}(z_i)-(1-\pi_d(z_i))\theta_{d0}(z_i)]^{1-\tilde{a}_{di}}.$$

For the validation sample, we can either select the m values at random (the 2S approach) or select $m=m_1+m_0$ values conditioned on the original observed realization of $\tilde{A}$ (the 3S approach). The validation sample portion of the likelihood under the 2S approach has the following components,

$$\begin{aligned} &P[A=a,\tilde{A}=\tilde{a}|D=d,Z=z_i] \\ &=P[\tilde{A}=\tilde{a}|A=a,D=d,Z=z_i]P[A=a|D=d,Z=z_i], \end{aligned}$$

whereas under the 3S approach,

$$\begin{aligned} &P[A=a|\tilde{A}=\tilde{a},D=d,Z=z_i] \\ &=\frac{P[\tilde{A}=\tilde{a}|A=a,D=d,Z=z_i]P[A=a|D=d,Z=z_i]}{P[\tilde{A}=\tilde{a}|D=d,Z=z_i]}, \end{aligned}$$

for $\tilde{a},a,d=0,1$ and $i=1,...,n$.

Thus, the validation sample likelihood for the 2S approach is,

$$\begin{aligned} L_V &= \prod_{\{v:v\in V\}}\theta_{d1}(z_v)\pi_d(z_v)^{a_{dv}\tilde{a}_{dv}}[\theta_{d0}(z_v)(1-\pi_d(z_v))]^{(1-a_{dv})\tilde{a}_{dv}}[(1-\theta_{d1}(z_v))\pi_d(z_v)]^{a_{dv}(1-\tilde{a}_{dv})} \\ &\quad\times[(1-\theta_{d0}(z_v))(1-\pi_d(z_v))]^{(1-a_{dv})(1-\tilde{a}_{dv})}, \end{aligned}$$

where V denotes the set of observations doubly measured by both the fallible and infallible classifiers. For the 3S approach, we are drawing two validation samples, one from either subset of the original EHR data with observed values of $\tilde{a} = 0, 1$. Thus, the likelihoods will be,

$$\begin{aligned} L_{V_1} &= \prod_{d=0,1} \prod_{\{v: v \in V, \tilde{a}_{dv} = 1\}} [\frac{\theta_{d1}(z_v)\pi_d(z_v)}{\pi_d(z_v)\theta_{d1}(z_v) + (1 - \pi_d(z_v))\theta_{d0}(z_v)}]^{a_{dv}\tilde{a}_{dv}} \\ &\quad \times [\frac{\theta_{d0}(z_v)(1 - \pi_d(z_v))}{\pi_d(z_v)\theta_{d1}(z_v) + (1 - \pi_d(z_v))\theta_{d0}(z_v)}]^{(1-a_{dv})\tilde{a}_{dv}} \end{aligned}$$

$$\begin{aligned} L_{V_0} &= \prod_{d=0,1} \prod_{\{v: v \in V, \tilde{a}_{dv} = 0\}} [\frac{(1 - \theta_{d1}(z_v))\pi_d(z_v)}{1 - \pi_d(z_v)\theta_{d1}(z_v) - (1 - \pi_d(z_v))\theta_{d0}(z_v)}]^{a_{dv}(1-\tilde{a}_{dv})} \\ &\quad \times [\frac{(1 - \theta_{d0}(z_v))(1 - \pi_d(z_v))}{1 - \pi_d(z_v)\theta_{d1}(z_v) - (1 - \pi_d(z_v))\theta_{d0}(z_v)}]^{(1-a_{dv})(1-\tilde{a}_{dv})}. \end{aligned}$$

Finally, the combined likelihoods are $L_2 = L_O \times L_V$ for the 2S approach and $L_3 = L_O \times L_{V_1} \times L_{V_0}$ for the 3S approach. Recall from Chapter 2, under 2S sampling, the joint probabilities account for the dependency between the two measurements and we must only consider the original realized outcome once in the likelihood. Hence, the second product in L_O will be taken over the set of observations, $\{i, d : O_d \cap V^c\}$, where O_d is used to denote all observations in the original EHR data with drug exposure status, d, and V^c is used to represent the set complement of V. The union of these sets with respect to $d = 0, 1$ will contain $n - m$ observations. Further, in this context, the likelihoods do not simplify in a similar fashion as in equation (2.6) and (2.8).

Next, we will consider modelling the π and θ-parameters in the context of the analysis of EHR data with the presence of binary outcome misclassification. Our primary concern in the analysis of this data is to estimate the association between drug utilization and AE occurrence. However, we are also interested in the relationship between AE status and other explanatory variables. Thus, we will model the π-parameters as,

$$\eta_{di} \;=\; logit(\pi_d(z_i)) = \alpha + \beta d + z_i^T \psi.$$

For the θ-parameters, we wish to model the misclassification rates so as to depend on a set of explanatory variables in addition to drug utilization status. This is desirable as it seems reasonable that misdiagnoses would likely depend on an overall patient profile. For instance, if a drug is suspected to be associated with another AE other than that under study, clinicians may be more likely to diagnose the alternate AE. To account for this we will define,

$$\rho_{dai} \;=\; logit(\theta_{da}(z_i)) = \kappa_{da} + z_i^T \tau.$$

Using these models we can characterize the log likelihoods and determine the MLEs via numerical maximization algorithms such as *nlminb*() or *nlm*() in R. For the 2S approach, the log likelihood based on the original data simplifies to

$$\begin{aligned} l_O \;=\; & \sum_{d=0,1} \sum_{\{i:O_d \cap V^c\}} (1 - \tilde{a}_{di}) \log(1 + e^{\eta_{di}} + e^{\rho_{d1i}} + e^{\eta_{di} + \rho_{d0i}}) \\ & + \tilde{a}_{di} \log(e^{\rho_{d0i}} + e^{\eta_{di} + \rho_{d1i}} + e^{\rho_{d1i} + \rho_{d0i}} + e^{\eta_{di} + \rho_{d1i} + \rho_{d0i}}) \\ & - \log(1 + e^{\eta_{di}}) - \log(1 + e^{\rho_{d1i}}) - \log(1 + e^{\rho_{d0i}}), \end{aligned}$$

and the validation sample log likelihood can be written as,

$$\begin{aligned} l_V \;=\; & \sum_{d=0,1} \sum_{\{v:v \in V\}} a_{dv}\eta_{dv} + a_{dv}\tilde{a}_{dv}\rho_{d1v} + \tilde{a}_{dv}(1 - a_{dv})\rho_{d0v} \\ & - \log(1 + e^{\eta_{dv}}) - a_{dv} \log(1 + e^{\rho_{d1v}}) - (1 - a_{dv}) \log(1 + e^{\rho_{d0v}}). \end{aligned}$$

The combined log likelihood is,

$$\begin{aligned} l_2 \;=\; & \sum_{d=0,1}\sum_{\{i:O_d\cap V^c\}}(1-\tilde{a}_{di})\log(1+e^{\eta_{di}}+e^{\rho_{d1i}}+e^{\eta_{di}+\rho_{d0i}}) \\ & +\tilde{a}_{di}\log(e^{\rho_{d0i}}+e^{\eta_{di}+\rho_{d1i}}+e^{\rho_{d1i}+\rho_{d0i}}+e^{\eta_{di}+\rho_{d1i}+\rho_{d0i}})-\log(1+e^{\eta_{di}})-\log(1+e^{\rho_{d1i}}) \\ & -\log(1+e^{\rho_{d0i}})+\sum_{\{v:v\in V\}}a_{dv}(1-\tilde{a}_{dv})\eta_{dv}+a_{dv}\tilde{a}_{dv}(\eta_{dv}+\rho_{d1v}) \\ & +\tilde{a}_{dv}(1-a_{dv})\rho_{d0v}-\log(1+e^{\eta_{dv}})-a_{dv}\log(1+e^{\rho_{d1v}})-(1-a_{dv})\log(1+e^{\rho_{d0v}}). \end{aligned}$$

For the 3S approach, l_O will be functionally identical to that given above but the summation will be over all observations in $O = O_1 \cup O_2$ and the validation sample will have the following log-likelihoods,

$$\begin{aligned} l_{V_1} \;=\; & \sum_{d=0,1}\sum_{\{v:v\in V,\tilde{a}_{dv}=1\}} a_{dv}\tilde{a}_{dv}(\eta_{dv}+\rho_{d1v})+\tilde{a}_{dv}(1-a_{dv})\rho_{d0v}+(1-a_{dv})\tilde{a}_{dv}\log(1+e^{\rho_{d1i}}) \\ & +a_{dv}\tilde{a}_{dv}\log(1+e^{\rho_{d0i}})-\tilde{a}_{dv}\log(e^{\rho_{d0v}}+e^{\eta_{dv}+\rho_{d1v}}+e^{\rho_{d1v}+\rho_{d0v}}+e^{\eta_{dv}+\rho_{d1v}+\rho_{d0v}}), \end{aligned}$$

$$\begin{aligned} l_{V_0} \;=\; & \sum_{d=0,1}\sum_{\{v:v\in V,\tilde{a}_{dv}=0\}} a_{dv}(1-\tilde{a}_{dv})\eta_{dv}+(1-a_{dv})(1-\tilde{a}_{dv})\log(1+e^{\rho_{d1i}}) \\ & +a_{dv}(1-\tilde{a}_{dv})\log(1+e^{\rho_{d0i}})-(1-\tilde{a}_{dv})\log(1+e^{\eta_{dv}}+e^{\rho_{d1v}}+e^{\eta_{dv}+\rho_{d0v}}). \end{aligned}$$

Note, that since the set of observations V is the union of two disjoint sets, namely those observations with original AE status $\tilde{A} = 1$ or 0, the combined log likelihood can be written as,

$$\begin{aligned} l_3 \;=\; & \sum_{d=0,1}\sum_{i=1}^{n}(1-\tilde{a}_{di})\log(1+e^{\eta_{di}}+e^{\rho_{d1i}}+e^{\eta_{di}+\rho_{d0i}})+\tilde{a}_{di}\log(e^{\rho_{d0i}}+e^{\eta_{di}+\rho_{d1i}}+e^{\rho_{d1i}+\rho_{d0i}}+e^{\eta_{di}+\rho_{d1i}+\rho_{d0i}}) \\ & -\log(1+e^{\eta_{di}})-\log(1+e^{\rho_{d1i}})-\log(1+e^{\rho_{d0i}})+\sum_{\{v:v\in V\}}a_{dv}\eta_{dv}+a_{dv}\tilde{a}_{dv}\rho_{d1v}+\tilde{a}_{dv}(1-a_{dv})\rho_{d0v} \\ & +a_{dv}\log(1+e^{\rho_{d0v}})+(1-a_{dv})\log(1+e^{\rho_{d1v}})-(1-\tilde{a}_{dv})\log(1+e^{\eta_{dv}}+e^{\rho_{d1v}}+e^{\eta_{dv}+\rho_{d0v}}) \\ & -\tilde{a}_{dv}\log(e^{\rho_{d0v}}+e^{\eta_{dv}+\rho_{d1v}}+e^{\rho_{d1v}+\rho_{d0v}}+e^{\eta_{dv}+\rho_{d1v}+\rho_{d0v}}). \end{aligned}$$

The use of a multi-sample framework to model these likelihoods allows us to use the inverse of the observed information matrix as a consistent estimator of the asymptotic variance under both 2S and 3S approaches. The underlying theoretical basis for this is outlined in detail in Appendix A; however, we will reproduce some of the important results here. The rationale for the use of this framework was presented in Section 2.3, and is equally applicable here. We are still modelling outcome misclassification with binary data, so that the discussion in Section 2.3 demonstrating that Definition A.2.1 is satisfied still holds. Thus, the observed information can be written as,

$$\frac{1}{n}\sum_{s=1}^{S} n_s \bar{I}_s(\hat{\gamma}) \xrightarrow{p} \sum_{s=1}^{S} \omega_s \mathcal{I}_s(\gamma_0), \tag{4.2}$$

where we have defined the p-dimensional parameter vector to be γ with MLE $\hat{\gamma}$, $\bar{I}_s(\hat{\gamma}) = \frac{1}{n_s}\sum_{i=1}^{n_s} I_{si}(\hat{\gamma}) = \frac{1}{n_s}\sum_{i=1}^{n_s}\left[-\frac{\partial l_s(\gamma|x_{si})}{\partial\gamma\partial\gamma^T}\right]_{\gamma=\hat{\gamma}}$, $n = \sum_{s=1}^{S} n_s$, and n_s is the size of the s^{th} sample. The limiting sum is the inverse of the asymptotic variance-covariance matrix, which is equivalent to the inverse of Fisher's information matrix and $\mathcal{I}_s(\gamma) = E[-\frac{\partial^2 l_s(\gamma)}{\partial\gamma\partial\gamma^T}]$ for the s^{th} sample. These results rely on the usual regularity conditions (see Appendix A) as well as convergence of the sample weights, $\frac{n_s}{n}$, to limiting weights that sum to one, $\left(\frac{n_1}{n}, ..., \frac{n_S}{n}\right) \longrightarrow (\omega_1, ..., \omega_s)$, $\sum_s \omega_s = 1$.

This allows us to write the following estimates of the asymptotic variance-covariance matrices under both the 2S and 3S approaches,

$$\hat{\sigma}^2(\hat{\gamma}_2) = ((n-m)\bar{I}_O(\hat{\gamma}) + m\bar{I}_V(\hat{\gamma}))^{-1}, \tag{4.3}$$

and,

$$\hat{\sigma}^2(\hat{\gamma}_3) = (n\bar{I}_O(\hat{\gamma}) + m_1\bar{I}_{V_1}(\hat{\gamma}) + m_0\bar{I}_{V_0}(\hat{\gamma}))^{-1} = (n\bar{I}_O(\hat{\gamma}) + m[P_1\bar{I}_{V_1}(\hat{\gamma}) + (1-P_1)\bar{I}_{V_0}(\hat{\gamma})])^{-1}. \tag{4.4}$$

The second expression in $\hat{\sigma}^2(\hat{\gamma}_3)$ allows us to write the validation sample components as a weighted sum with weights $P_1 = \frac{m_1}{m}$, which denotes the proportion of category 1 observations in the validation sample. Thus, we are able to observe $\hat{\sigma}^2(\hat{\gamma}_3)$ as P_1 varies in a similar manner to Chapter 2. In Appendix C, R code is presented to compute the elements of the $\bar{I}_s$, $s = 1, ..., S$, matrices. In the next section we can use these formulas in simulation studies to numerically examine the properties of these estimators as well as consider the ARE between the two sampling approaches.

4.3 Comparative Simulation Study for Logistic Regression: 2S versus 3S Approach

In this section, we conduct simulation studies designed to investigate these methods of estimation in more detail. We will model the π and θ-parameters as function of both binary and continuous covariates and simulate from an extensive list of target parameters chosen to replicate EHR datasets. Thus, we will specify two additional covariates in addition to drug utilization status that are realistic with respect to our example. The first will be an additional binary covariate such as gender, for example. The second will be a positive continuous covariate, such as a measure of a biochemical variable related to AE occurrence. To incorporate these additional variables into our data generation algorithm we will generate these observations from a binomial random generator with success probability 0.6 and a gamma random number generator with parameters (2,1).

To select the simulation target parameters, first recall that we must be cautious about generating scenarios in which sample size issues may be encountered. We would ideally like to consider values of $\gamma = (\alpha, \beta, \psi_1, \psi_2, \kappa_{11}, \kappa_{01}, \kappa_{10}, \kappa_{00}, \tau_1, \tau_2)$ that are likely to produce reasonably large sample sizes within all groups. However, in reality, small rates of misclassification coupled with rare events are characteristics of EHR data, a situation that may cause difficulties in estimation. To emulate these situations we will select sets of values that mimic these extreme examples, keeping in mind that the success of these methods in producing MCB-adjusted estimators will likely only improve as more data is accessible in the validation subgroups. Additional simulation scenarios beyond those presented here are included in Appendix D.

Recall that we are modelling the logits of the π and θ-parameters as,

$$\begin{aligned}
\eta_{di} &= \alpha + \beta d + \psi_1 z_{i1} + \psi_2 z_{i2} \\
\rho_{11i} &= \kappa_{11} + \tau_1 z_{i1} + \tau_2 z_{i2} \\
\rho_{01i} &= \kappa_{01} + \tau_1 z_{i1} + \tau_2 z_{i2} \\
\rho_{10i} &= \kappa_{10} + \tau_1 z_{i1} + \tau_2 z_{i2} \\
\rho_{00i} &= \kappa_{00} + \tau_1 z_{i1} + \tau_2 z_{i2}.
\end{aligned}$$

For the η-parameters, we are primarily interested in testing $H_0 : \beta = 0$, since in the motivating example, our primary interest is in the association between AE occurrence and drug utilization status. Since the intercept α can be thought of as the log-odds ratio of the frequency of AE occurrence among patients not taking the drug, this allows us to select some realistic baseline values and then select the β's in a manner that introduces an enhanced (or diminished) association, taking into account the effects of the additional covariates on the outcome of interest. In other words, we can start by considering realistic values for the actual π-parameters and iteratively solve the target regression parameters. Thus,

$$\alpha = \log\left(\frac{\pi_0(0)}{1-\pi_0(0)}\right), \tag{4.5}$$

where $\pi_d(z_1) = P(A = 1|D = d, Z_1 = z_1)$ for $d, z_1 = 0$. The same probability in the $d = 1$ group is defined by,

$$\beta = \log\left(\frac{\pi_1(0)}{1-\pi_1(0)}\right) - \alpha. \tag{4.6}$$

For the coefficient to the additional binary covariate, we can specify this in the same manner as β,

$$\psi_1 = \log\left(\frac{\pi_1(1)}{1-\pi_1(1)}\right) - \alpha - \beta. \tag{4.7}$$

Finally, ψ_2 is the coefficient to a vector of observations with a positive continuous distribution. To incorporate this information into the model, we will utilize the mean of the distribution used to generate the values of z_2. Thus,

$$\psi_2\mu_{z_2} = \log\left(\frac{\pi_1(1,\mu_{z_2})}{1-\pi_1(1,\mu_{z_2})}\right) - \alpha - \beta - \psi_1, \tag{4.8}$$

where $\pi_d(z_1, z_2) = P(A = 1|D = d, Z_1 = z_1, Z_2 = z_2)$. Since we chose a gamma(2,1) to generate z_2 in this set of simulation studies, $\mu_{z_2} = 2$, $\psi_2 = \frac{1}{2}\left[\log\left(\frac{\pi_1(1,\mu_{z_2})}{1-\pi_1(1,\mu_{z_2})}\right) - \alpha - \beta - \psi_1\right]$.

Next, to generate the misclassification probability regression parameters based on specification of the underlying probabilities, we must consider selection differently. For each d, a pair, we model z_1 and z_2 using the same coefficient. This implies that the differences in the logit of the θ-parameter in each group are being held constant. In other words,

$$\tau_1 = \log\left(\frac{\theta_{da}(1)}{1-\theta_{da}(1)}\right) - \kappa_{da}, \tag{4.9}$$

where $\theta_{da}(z_1) = P[\tilde{A} = 1|D = d, A = a, Z_1 = z_1]$ and $\kappa_{da} = \log\left(\frac{\theta_{da}(0)}{1-\theta_{da}(0)}\right)$, $d, a = 0, 1$. Thus, we can first specify the $\theta_{da}(0)$'s followed by the difference, which forces the values of $\theta_{da}(1)$, however, this makes sense given the context of the modelling. For example, consider that gender has an effect on the misclassification rate, which is a reasonable hypothesis over a range of diverse drug/AE pairs. This specification asserts that the log odds-ratio with respect to gender is the same regardless of the drug utilization group or even the AE outcome group to which an individual belongs. Thus, the log odds-ratio for the presence or absence of z_1 is fixed at τ_1 for all $d, a = 0, 1$.

For τ_2, we can apply similar reasoning in that we have solved the $\theta_{da}(1)$'s given selection of the difference, τ_1 allowing us to solve the next difference,

$$\tau_2 = \frac{1}{\mu_{z_2}} \left[\log \left(\frac{\theta_{da}(1, \mu_{z_2})}{1 - \theta_{da}(1, \mu_{z_2})} \right) - \tau_1 - \kappa_{da} \right], \tag{4.10}$$

where $\theta_{da}(z_1, z_2) = P[\tilde{A} = 1 | D = d, A = a, Z_1 = z_1, Z_2 = z_2]$ and $\mu_{z_2} = 2$ in our example.

Using this rationale we can select realistic simulation parameters as displayed in Table 4.3.3, by first selecting probabilities or log odds-ratios chosen to emulate realistic EHR scenarios which are displayed in Tables 4.3.1 and 4.3.2. We will return to this discussion in Section 4.4 when addressing the sample size determination issues.

Finally, recall our concerns regarding sample sizes that may be too small in the subgroups of the validation sample. This will influence our decision regarding realistic simulation parameters; however, we will also build in an error indicator that will specify if any results produced validation sample sizes that fall short of a minimal threshold. To document any trends in the size of the standard error in estimation with respect to different values of P_1, we will begin simulating data with respect to the 3S approach over a range of P_1 values, $P_1 \in (0.1, 0.2, ..., 0.9)$.

For the overall validation sample size, n, we will select 10,000 given that these datasets are generally quite large and investigate increasing values of m. Recall that we are attempting to mimic extreme examples and, as such, a 10% validation sample size was inadequate across most simulations to draw sufficiently large subgroup sample sizes; hence, we will consider validation sample sizes of 20% and 30% to begin. Note, that the 10,000 observations will be first divided into even drug utilization groups with 5000 in category 1 and 5000 in category 2. To characterize the impact as this ratio changes, we will also consider 30%/70% and 70%/30% breakdowns.

Finally, returning to our motivating example, we will consider scenarios from which to generate our target simulation parameters. To begin, consider the π-parameters. The first scenario, will presume a small probability of AE occurrence in the $d = 0$ group which increases with the presence of the drug. In this scenario, the additional covariates will have no impact on the probability of AE occurrence. Next, we will introduce a small increase in AE occurrence rate for each of the additional covariates. For our final scenario, we will remove the increase due to drug presence but introduce a large increase in the AE rates due to the presence of z_1. These are presented in Table 4.3.1, where each scenario is marked by the column heading.

Table 4.3.1: Chosen target π-parameters for the simulation study comparing the 2S and 3S approach for logistic regression.

Parameter	1	2	3
$\pi_0(0,0)$	0.1	0.1	0.1
$\pi_1(0,0)$	0.25	0.25	0.1
$\pi_1(1,0)$	0.25	0.28	0.35
$\pi_1(1,\mu_{z_2})$	0.25	0.3	0.35

For the κ-parameters, we will first introduce misclassification with the same rate for each of the drug groups. Following this, we will increase the rate only in the $d = 1$ group, mimicking a situation in which a clinician may be suspicious of an adverse drug reaction. For the τ-parameters we will consider odds-ratios of 1 for both of the additional covariates. In other words, the presence or absence of z_1 and the effect of a mean increase in z_2 will not alter the misclassification rates.

However, in the next scenario we will implement odds-ratios greater than 1 for both z_1 and z_2. These four scenarios are described in Table 4.3.2.

Table 4.3.2: Chosen target θ and τ-parameters for the simulation study comparing the 2S and 3S approach for logistic regression.

Parameter	1	2	3	4
$\theta_{11}(0,0)$	0.9	0.9	0.9	0.9
$\theta_{01}(0,0)$	0.92	0.92	0.92	0.92
$\theta_{10}(0,0)$	0.1	0.15	0.1	0.15
$\theta_{00}(0,0)$	0.08	0.05	0.08	0.05
τ_1	0	0	0.1	0.1
τ_2	0	0	0.05	0.05

We thus have the twelve scenarios summarized in Table 4.3.3. Scenario 1 is π-parameter 1 and θ-parameter 1, scenario 2 is π-parameter 1 and θ-parameter 2, and so on.

Table 4.3.3: Target regression parameters for the simulation study comparing the 2S and 3S approach for logistic regression.

Scenario	α	β	ψ_1	ψ_2	κ_{11}	κ_{10}	κ_{01}	κ_{00}	τ_1	τ_2
1	-2.197	1.099	0	0	2.197	2.442	-2.197	-2.442	0	0
2	-2.197	1.099	0	0	2.197	2.442	-1.735	-2.944	0	0
3	-2.197	1.099	0	0	2.197	2.442	-2.197	-2.442	0.1	0.05
4	-2.197	1.099	0	0	2.197	2.442	-1.735	-2.944	0.1	0.05
5	-2.197	1.099	0.154	0.049	2.197	2.442	-2.197	-2.442	0	0
6	-2.197	1.099	0.154	0.049	2.197	2.442	-1.735	-2.944	0	0
7	-2.197	1.099	0.154	0.049	2.197	2.442	-2.197	-2.442	0.1	0.05
8	-2.197	1.099	0.154	0.049	2.197	2.442	-1.735	-2.944	0.1	0.05
9	-2.197	0	1.578	0	2.197	2.442	-2.197	-2.442	0	0
10	-2.197	0	1.578	0	2.197	2.442	-1.735	-2.944	0	0
11	-2.197	0	1.578	0	2.197	2.442	-2.197	-2.442	0.1	0.05
12	-2.197	0	1.578	0	2.197	2.442	-1.735	-2.944	0.1	0.05

These parameters are chosen to represent realistic scenarios for EHR data, as well as to demonstrate the problematic effect of validation subsample sizes sufficiently small that problems in estimation resulting in bias may be encountered. Other simulations were conducted under additional scenarios and the results in Appendix D indicate that both methods both perform well given adequate sample sizes. As a consequence, we will confine the discussion to the parameter sets in Table 4.3.3. Results for selected simulation scenarios are provided in Tables 4.3.4 - 4.3.8.

Table 4.3.4: Results for parameter set 1 comparing the 2S and 3S approach for logistic regression with n=10,000, m=2,000.

(a) Summary Table

Parameter	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\hat{\gamma}_2$	$\hat{\sigma}(\hat{\gamma}_2)$	s	$\hat{\gamma}_3^{in}$	$\hat{\sigma}(\hat{\gamma}_3^{in})$	s	$\hat{\gamma}_3^{min}$	$\hat{\sigma}(\hat{\gamma}_3^{min})$	s
α	-2.197	-1.6259	0.0554	-2.2004	0.1122	0.1127	-2.1926	0.1175	0.1187	-2.1967	0.0964	0.0985
β	1.099	0.7781	0.0491	1.098	0.1	0.103	1.0938	0.1044	0.1086	1.0961	0.0872	0.0866
ψ_1	0	0.0001	0.049	0.0026	0.0917	0.0906	-0.0004	0.0949	0.0917	-0.0004	0.0787	0.0775
ψ_2	0	-0.001	0.0182	-0.0002	0.0346	0.0332	-0.0026	0.0346	0.0346	-0.0008	0.03	0.03
κ_{11}	2.197	-	-	2.2186	0.2296	0.2435	2.224	0.2289	0.2371	2.2268	0.2482	0.2528
κ_{01}	2.442	-	-	2.5243	0.4055	0.4251	2.5167	0.3961	0.4044	2.5572	0.4497	0.444
κ_{10}	-2.197	-	-	-2.1981	0.1463	0.151	-2.2085	0.1552	0.1597	-2.2001	0.12	0.1187
κ_{00}	-2.442	-	-	-2.4416	0.1404	0.14	-2.4538	0.1483	0.1503	-2.4425	0.1166	0.1166
τ_1	0	-	-	-0.0022	0.1241	0.1269	0.0024	0.1296	0.1269	0.0013	0.1063	0.1063
τ_1	0	-	-	-0.0035	0.0458	0.0458	-0.0006	0.048	0.05	-0.0021	0.0387	0.0387

(b) Coverage Proportion Estimates.

Parameter	$\widetilde{CP}$	$\widehat{CP}_2$	$\widehat{CP}_3^{in}$	$\widehat{CP}_3^{min}$
α	0	0.951	0.946	0.94
β	0	0.946	0.931	0.956
ψ_1	0.946	0.944	0.965	0.955
ψ_2	0.952	0.95	0.959	0.953
κ_{11}	-	0.944	0.944	0.95
κ_{01}	-	0.962	0.965	0.969
κ_{10}	-	0.946	0.939	0.951
κ_{00}	-	0.95	0.931	0.951
τ_1	-	0.942	0.958	0.95
τ_1	-	0.953	0.942	0.952

Table 4.3.5: Results for parameter set 1 comparing the 2S and 3S approach for logistic regression with n=10,000, m=3,000.

(a) Summary Table

Parameter	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\hat{\gamma}_2$	$\hat{\sigma}(\hat{\gamma}_2)$	s	$\hat{\gamma}_3^{in}$	$\hat{\sigma}(\hat{\gamma}_3^{in})$	s	$\hat{\gamma}_3^{min}$	$\hat{\sigma}(\hat{\gamma}_3^{min})$	s
α	-2.197	-1.6255	0.0554	-2.1973	0.0959	0.102	-2.1968	0.1	0.1044	-2.1915	0.08	0.0825
β	1.099	0.7831	0.0491	1.1013	0.0849	0.0872	1.1022	0.0889	0.0837	1.0984	0.0735	0.0742
ψ_1	0	-0.0015	0.049	-0.0005	0.0787	0.0843	0.0002	0.0812	0.0849	-0.0053	0.0648	0.0686
ψ_2	0	-0.0012	0.0182	-0.002	0.0283	0.0283	-0.0022	0.03	0.03	-0.0007	0.0245	0.0245
κ_{11}	2.197	-	-	2.2213	0.1892	0.1849	2.2129	0.1876	0.1975	2.2258	0.2184	0.2173
κ_{01}	2.442	-	-	2.4919	0.3211	0.3219	2.4902	0.3165	0.3197	2.5115	0.3969	0.4164
κ_{10}	-2.197	-	-	-2.1963	0.1233	0.1208	-2.1972	0.1296	0.1353	-2.1994	0.0954	0.0964
κ_{00}	-2.442	-	-	-2.4424	0.1192	0.1136	-2.4443	0.1253	0.1249	-2.4462	0.0943	0.0954
τ_1	0	-	-	-0.0018	0.1058	0.102	-0.0021	0.11	0.1131	0.0047	0.086	0.0866
τ_1	0	-	-	-0.0009	0.0387	0.0374	-0.0007	0.04	0.04	-0.0023	0.0316	0.03

(b) Coverage Proportion Estimates.

Parameter	$\widehat{CP}$	$\widehat{CP}_2$	$\widehat{CP}_3^{in}$	$\widehat{CP}_3^{min}$
α	0	0.92	0.942	0.945
β	0	0.95	0.967	0.948
ψ_1	0.925	0.937	0.941	0.934
ψ_2	0.945	0.956	0.948	0.934
κ_{11}	-	0.946	0.941	0.951
κ_{01}	-	0.962	0.969	0.953
κ_{10}	-	0.957	0.941	0.933
κ_{00}	-	0.961	0.948	0.948
τ_1	-	0.967	0.937	0.964
τ_1	-	0.958	0.948	0.957

Table 4.3.6: Results for parameter set 3 comparing the 2S and 3S approach for logistic regression with n=10,000, m=2,000.

(a) Summary Table

Parameter	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\hat{\gamma}_2$	$\hat{\sigma}(\hat{\gamma}_2)$	s	$\hat{\gamma}_3^{in}$	$\hat{\sigma}(\hat{\gamma}_3^{in})$	s	$\hat{\gamma}_3^{min}$	$\hat{\sigma}(\hat{\gamma}_3^{min})$	s
α	-2.197	-1.6197	0.0546	-2.1984	0.1136	0.1127	-2.2029	0.1212	0.1204	-2.207	0.1063	0.102
β	1.099	0.7659	0.0482	1.1028	0.101	0.1015	1.1017	0.1082	0.1114	1.107	0.0949	0.0954
ψ_1	0	0.0496	0.0483	-0.0021	0.0917	0.0933	0.0035	0.097	0.0954	0.0031	0.0866	0.0837
ψ_2	0	0.0245	0.0176	-0.0009	0.0346	0.0346	-0.0009	0.0361	0.0374	0.001	0.0316	0.0316
κ_{11}	2.197	-	-	2.2267	0.2371	0.2394	2.2292	0.2341	0.2328	2.2302	0.2423	0.2506
κ_{01}	2.442	-	-	2.5587	0.4288	0.4498	2.5261	0.4146	0.4146	2.5467	0.4478	0.4571
κ_{10}	-2.197	-	-	-2.2088	0.1404	0.1439	-2.2007	0.1503	0.1513	-2.2002	0.1296	0.1285
κ_{00}	-2.442	-	-	-2.4488	0.1353	0.1386	-2.4449	0.1449	0.1446	-2.4366	0.1249	0.1245
τ_1	0.1	-	-	0.106	0.1192	0.1233	0.0985	0.1253	0.1277	0.0984	0.1118	0.1082
τ_1	0.05	-	-	0.0484	0.0412	0.0424	0.0483	0.0436	0.0436	0.0462	0.0387	0.04

(b) Coverage Proportion Estimates.

Parameter	$\widehat{CP}$	$\widehat{CP}_2$	$\widehat{CP}_3^{in}$	$\widehat{CP}_3^{min}$
α	0	0.948	0.954	0.952
β	0	0.949	0.938	0.945
ψ_1	0.813	0.945	0.959	0.958
ψ_2	0.705	0.947	0.941	0.944
κ_{11}	-	0.955	0.963	0.945
κ_{01}	-	0.961	0.973	0.973
κ_{10}	-	0.943	0.95	0.949
κ_{00}	-	0.947	0.955	0.951
τ_1	-	0.943	0.941	0.958
τ_1	-	0.947	0.944	0.944

Table 4.3.7: Results for parameter set 3 comparing the 2S and 3S approach for logistic regression with n=10,000, m=3,000.

(a) Summary Table

Parameter	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\hat{\gamma}_2$	$\hat{\sigma}(\hat{\gamma}_2)$	s	$\hat{\gamma}_3^{in}$	$\hat{\sigma}(\hat{\gamma}_3^{in})$	s	$\hat{\gamma}_3^{min}$	$\hat{\sigma}(\hat{\gamma}_3^{min})$	s
α	-2.197	-1.6173	0.0546	-2.1824	0.0964	0.11	-2.1961	0.103	0.1034	-2.196	0.0806	0.0906
β	1.099	0.765	0.0482	1.0925	0.0854	0.0877	1.0963	0.0911	0.0883	1.1022	0.0735	0.0775
ψ_1	0	0.0465	0.0483	-0.0094	0.0787	0.0843	-0.0026	0.0831	0.086	-0.003	0.0656	0.0714
ψ_2	0	0.0237	0.0177	-0.003	0.03	0.03	-0.0001	0.0316	0.03	-0.0011	0.0245	0.0245
κ_{11}	2.197	-	-	2.2103	0.1942	0.1803	2.2113	0.1913	0.1897	2.2176	0.2263	0.2247
κ_{01}	2.442	-	-	2.4965	0.3394	0.3704	2.4763	0.3276	0.3366	2.5133	0.4175	0.4338
κ_{10}	-2.197	-	-	-2.2114	0.1187	0.1245	-2.1933	0.1265	0.1249	-2.2013	0.0933	0.0922
κ_{00}	-2.442	-	-	-2.4616	0.1158	0.1257	-2.4406	0.1225	0.1179	-2.4401	0.0917	0.0889
τ_1	0.1	-	-	0.1095	0.102	0.1	0.0988	0.1068	0.1054	0.1	0.0837	0.0849
τ_1	0.05	-	-	0.0505	0.0361	0.0361	0.0463	0.0374	0.0374	0.0482	0.03	0.0283

(b) Coverage Proportion Estimates.

Parameter	$\widetilde{CP}$	$\widehat{CP}_2$	$\widehat{CP}_3^{in}$	$\widehat{CP}_3^{min}$
α	0	0.926	0.955	0.918
β	0	0.938	0.958	0.936
ψ_1	0.829	0.942	0.939	0.928
ψ_2	0.735	0.941	0.947	0.958
κ_{11}	-	0.969	0.962	0.957
κ_{01}	-	0.935	0.954	0.96
κ_{10}	-	0.957	0.947	0.943
κ_{00}	-	0.937	0.959	0.961
τ_1	-	0.951	0.946	0.943
τ_1	-	0.948	0.947	0.964

For each presented parameter set, we display a table with two panels, (a) and (b). Panel (a) presents the point estimates and estimates of the standard error for the standard methodology ignoring outcome misclassification. The column marked $\tilde{\gamma}$ was generated using the function *glm*() in R while the 2S approach is presented in the column marked $\hat{\gamma}_2$, the 3S approach with P_1 close to the incidence rate of category 1 observations in the original data is presented in the column marked $\hat{\gamma}_3^{in}$, and using the value of $P_1 \in (0.1, ..., 0.9)$ that minimizes the variance, $\hat{\gamma}_3^{min}$. For the estimates of standard error, the columns marked with a $\hat{\sigma}$ are those calculated from the inverse of the observed information matrix while the columns marked 's' denote the sample standard deviation of the generated MLEs. Finally, the first column marked 'Target' denotes the target parameter values as specified in Table 4.3.3. Next, panel (b) presents the coverage proportion estimates for each approach. Again the ‘tilde’ denotes the method ignoring misclassification and the ‘hat’ denotes the use of the MCB-adjustment methods discussed in this chapter where the subscripts and superscripts denote the specific adjustment approach.

The results in Tables 4.3.6 and 4.3.7 clearly indicate that the MCB-adjustment methods all perform quite well, and demonstrate accurate coverage probabilities. However, for the scenario considered in Table 4.3.7b, the 2S approach leads to slightly understated coverage for the κ_{00} parameter. Increasing the validation sample size, m to 50% (see Tables 4.3.8a and 4.3.8b) results in improved coverage estimates.

Table 4.3.8: Results for parameter set 3 comparing the 2S and 3S approach for logistic regression with n=10,000, m=5,000.

(a) Summary Table

Parameter	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\hat{\gamma}_2$	$\hat{\sigma}(\hat{\gamma}_2)$	s	$\hat{\gamma}_3^{in}$	$\hat{\sigma}(\hat{\gamma}_3^{in})$	s	$\hat{\gamma}_3^{min}$	$\hat{\sigma}(\hat{\gamma}_3^{min})$	s
α	-2.197	-1.622	0.0546	-2.2002	0.08	0.0775	-2.1961	0.0843	0.0825	-2.1992	0.0714	0.07
β	1.099	0.7672	0.0482	1.1044	0.0707	0.07	1.0996	0.0748	0.0742	1.1007	0.064	0.0616
ψ_1	0	0.0488	0.0483	-0.0025	0.0663	0.064	-0.0053	0.0693	0.0678	-0.0023	0.0592	0.0574
ψ_2	0	0.0251	0.0176	-0.0012	0.0245	0.0245	-0.0006	0.0265	0.0245	0.0001	0.0224	0.0224
κ_{11}	2.197	-	-	2.2061	0.153	0.1526	2.2058	0.1513	0.1497	2.21	0.1646	0.1619
κ_{01}	2.442	-	-	2.4796	0.2612	0.2757	2.4703	0.2544	0.2587	2.4874	0.2895	0.3087
κ_{10}	-2.197	-	-	-2.2039	0.0975	0.0949	-2.2031	0.1039	0.1015	-2.1989	0.0837	0.0837
κ_{00}	-2.442	-	-	-2.4461	0.0954	0.0938	-2.4509	0.101	0.1	-2.4454	0.0837	0.0837
τ_1	0.1	-	-	0.1035	0.0849	0.0849	0.1073	0.0889	0.0883	0.103	0.0755	0.0748
τ_1	0.05	-	-	0.0508	0.03	0.03	0.05	0.0316	0.0316	0.0495	0.0265	0.0265

(b) Coverage Proportion Estimates.

Parameter	$\widetilde{CP}$	$\widehat{CP}_2$	$\widehat{CP}_3^{in}$	$\widehat{CP}_3^{min}$
α	0	0.953	0.957	0.954
β	0	0.945	0.949	0.953
ψ_1	0.832	0.96	0.952	0.963
ψ_2	0.711	0.962	0.953	0.951
κ_{11}	-	0.949	0.953	0.959
κ_{01}	-	0.958	0.949	0.956
κ_{10}	-	0.955	0.958	0.958
κ_{00}	-	0.961	0.961	0.953
τ_1	-	0.944	0.948	0.952
τ_1	-	0.961	0.957	0.953

Consider next the standard errors of $\hat{\beta}$. Since β is the parameter of interest, we would like to select the approach that produces the smallest estimates of variance.

Table 4.3.4a displays the estimates for parameter set 1 calculated using the square root of diagonal elements of the inverse of the observed information matrix given in equations (4.3) and (4.4). Note, that the smallest estimate of the standard error for the parameter of interest, β, is in the column, $\hat{\sigma}(\hat{\gamma}_3^{min})$, at .0872 which is slightly lower than that for the 2S approach. Note that the associated value of P_1 is 0.4 which can be interpreted as a validation sample drawn with subsample sizes $m_1 = 800$ and $m_0 = 1200$. Note, that the Monte Carlo estimate of the category 1 incidence rate, $\bar{n}_1^{MC}/n$, is approximately 23.22%. For each iteration of the simulation, the 2S approach is likely to draw an m_1 value within the range $m\frac{n_1}{n} \pm z_{\alpha/2}\sqrt{\frac{n_1 n_0}{n^2 m}}$, where n_1 is the number of category 1 observations in the generated original sample. Thus, it is likely that the observed category 1 sample size in the 2S approach will be close to 232. Note that the estimate of the standard error associated with β in the $\hat{\gamma}_3^{in}$ column corresponds to $P_1 = 0.2$, which appears to demonstrate a decreasing trend. However, we have only considered three points of one simulation. Thus, for select parameter sets we will plot the 3S approach's estimates of standard error for $\hat{\beta}$ across the values of P_1 and will mark the 2S approach with a horizontal line. Recall that we ignore validation sampling schemes that produce subsample sizes that are extremely low in these simulations, and as such we will select an overall validation sample size of $m = 5000$, to hopefully minimize the occurrence of such simulation outcomes. Plots for all twelve parameter sets are presented in Figures 4.3.1-4.3.3.

Figure 4.3.1: Plot of Monte Carlo Estimates of $\hat{\sigma}(\hat{\beta})$ versus P_1 for parameter sets 1 to 4.

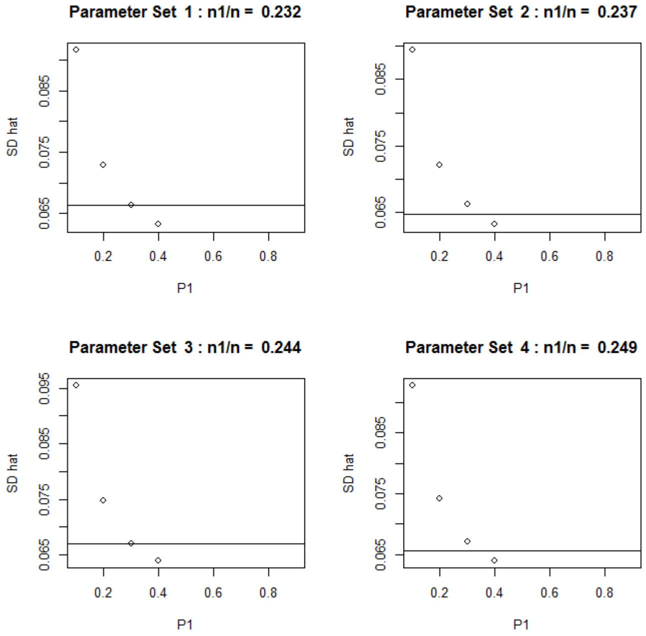


Figure 4.3.2: Plot of Monte Carlo Estimates of $\hat{\sigma}(\hat{\beta})$ versus P_1 for parameter sets 5 to 8.

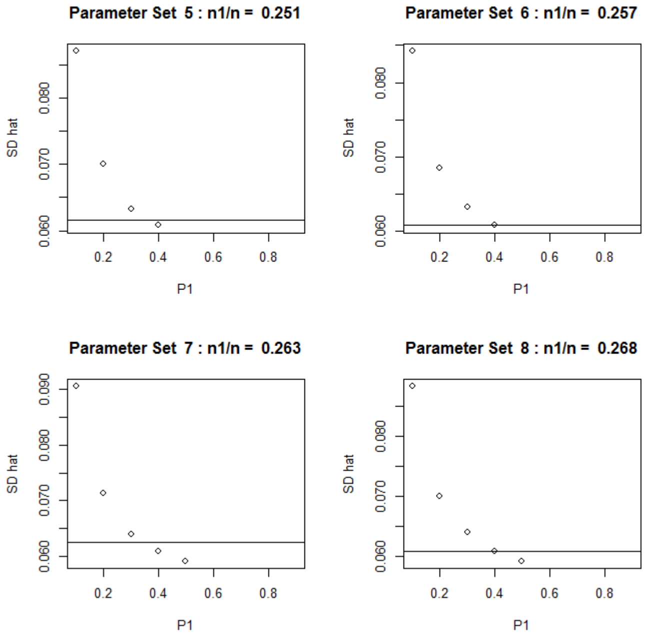
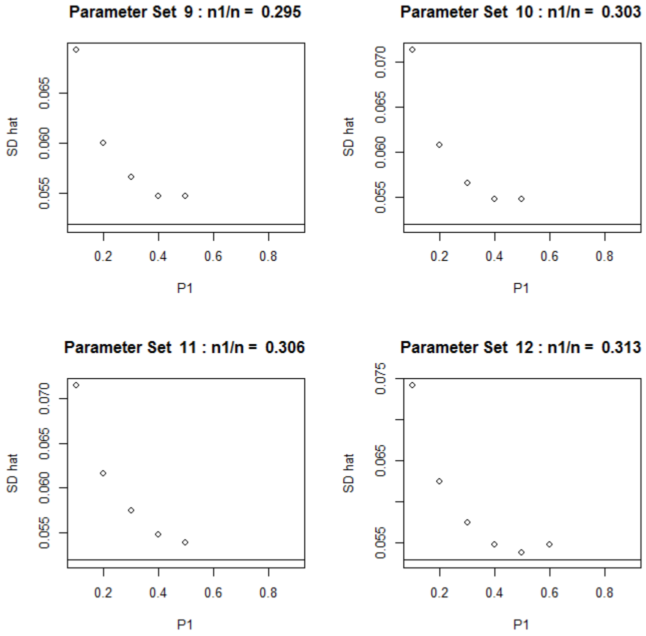


Figure 4.3.3: Plot of Monte Carlo Estimates of $\hat{\sigma}(\hat{\beta})$ versus P_1 for parameter sets 9 to 12.

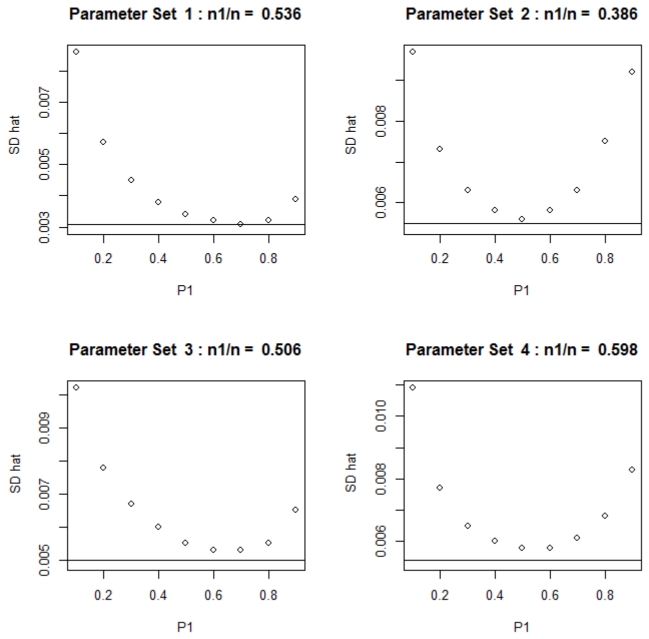


For the first set of four plots in Figure 4.3.1, the estimate of variance associated with the 3S approach is smaller than the 2S approach for $P_1 = 0.4$. Results for $P_1 > 0.4$ are not presented as they are drawing subsample sizes that are too low in this region. The second set of plots associated with parameter sets 5 - 8 demonstrate similar trends; however, for parameter sets 7 and 8, the 3S approach appears to be ideal when P1 = 0.5. In all other cases, the 2S approach appears to outperform the 3S approach across all allowable values of P_1. In the 3S approach, there appears to be a generally decreasing relationship between P_1 and $\hat{\sigma}(\hat{\beta})$ across all parameter sets; however, for parameter set 12, $\hat{\sigma}(\hat{\beta})$ begins to increase after achieving its minimal value at $P_1 = 0.5$. Recall, that in Chapter 2 we demonstrated that there exists a convex relationship between P_1 and $\hat{\sigma}(\hat{\gamma}_3)$ for a given parameter of interest. This was done using positive definite matrices in general and as such this result will continue to hold throughout this discussion. Up to this point, we have only considered situations in which extreme target simulation parameters were used and have chosen to ignore those P_1 values in which insufficient sample sizes are realized within the validation subgroups, which explains the lack of data points in Figures 4.3.1-4.3.3 for values of $P_1 > 0.5$. We now consider parameter sets that are less extreme, presented in Table 4.3.9, and produce similar plots to observe the relationship between P_1 and $\hat{\sigma}(\hat{\beta})$ across the entire set of P_1 values. These simulations were conducted with $m = 3000$.

Table 4.3.9: Chosen target regression parameters for additional simulation study comparing the 2S and 3S approach for logistic regression.

ID	α	β	ψ_1	ψ_2	κ_{11}	κ_{10}	κ_{01}	κ_{00}	τ_1	τ_2
1	-1	0	0.5	0.3	2	3	-3	-2	0	0.7
2	-0.5	2	-1	-0.5	2	2	-2	-2	0	0
3	-0.5	2	-1	-0.5	2	3	-3	-2	0	0.7
4	0	-2	1	0.8	2	2	-2	-2	0	0

Figure 4.3.4: Plot of Monte Carlo Estimates of $\hat{\sigma}(\hat{\beta})$ versus P_1 for additional target parameters from Table 4.3.9.



In these plots, we are able to observe the relationship between P_1 and $\hat{\sigma}(\hat{\beta})$ under the 3S approach across all values of $P_1 = (0.1, ..., 0.9)$. In all four plots, we see a parabolic looking function which achieves a minimum within the range of values of P_1 considered. In all cases, the 2S approach outperforms the 3S approach.

Note that the MC estimate of the observed incidence rates are much larger than in the previous scenarios reflected in Figures 4.3.1 - 4.3.3.

The results presented in this section demonstrate that the relationship between the 2S and 3S approach is not as easily characterized as in the contingency table methods discussed in Chapter 2. Thus, there is a larger need for a validation sample size determination algorithm in this section. However, immediate challenges in designing such an algorithm are due to the unknown and changing structure of modelling the additional covariates across applications, along with the manner in which they influence the underlying probabilistic structure of the data. First, experimenter-selected values based on educated guesses or information supplied via a pilot study on the assumed nature of the true regression coefficients is necessary, as opposed to the classification probabilities as in Chapter 2. This may be a complicated request as the inclination is to return an educated guess based on probabilities of occurrence. However, in building the simulation studies discussed throughout this section, we described a method for choosing the simulation regression parameters based on specification of probabilities as seen in Tables 4.3.1 and 4.3.2. Hence, a similar method can be used in practice to move from probabilities to regression coefficient specifications for a validation sample size determination algorithm. Next, the example used for the simulation studies is fairly simple in that we only consider three covariates; two of them binary and the third continuous and positive. There are many desired scenarios that will be difficult to apply to such an algorithm, either due to complexities in ascertaining the Monte Carlo estimates, or due to computational time for the simulation step in Algorithm 1 described in Section 3.3.

Thus, for this discussion, we will extend this algorithm to allow for the models discussed in Section 4.3 and demonstrate that they work reasonably well in approximating the resulting relationship between P_1, m and a bound on Bonferonni intervals based on the 2S and 3S approaches.

4.4 Monte Carlo Sample Size Determination Algorithm with Extensions for Logistic Regression

The data generation step of Algorithm 1 in Chapter 3 is used to produce approximations of the limiting observed information matrices for each of the s samples at the point of the parameter values specified by the investigators. This step is justified in Chapter 3, as we demonstrated that the observed information matrices converged in m as $m \longrightarrow n$. We further demonstrated that the matrices stabilize fairly quickly and recommended generating data at a value of $m = 0.3 \times n$. Next, we pointed out that for the 3S approach, there is also the potential for the chosen value of P_1 influencing the convergence behaviour. However, we demonstrated numerically that the difference between the approximations based on different chosen values of P_1 will converge to zero. While these properties were only investigated numerically for the contingency table approaches, we assert that they will continue to hold for the models discussed here.

For the contingency table approaches discussed in Chapter 3, the observed information matrices for the validation sample are a function of the investigator specified parameters (through an educated guess or pilot study) and the realized sample counts. Thus, convergence depends entirely on the influence of m and P_1 over these counts, which follow binomial and multinomial distributions. As such, categorizing the observations into subgroups should produce predictable expected counts regardless of the size of m other than for values so small that there are sample size issues. In the logistic regression case, we can employ similar reasoning since the additional covariate information will only increase the number of possible subgroups. Given that we can assume the availability of investigator specified parameters, the counts will again tend to their expectation and we should see similar behaviour with regards to convergence.

In this section, we will apply Algorithm 1 to the 2S and 3S approaches for logistic regression in a similar manner as was done for the contingency table approaches. Since we are modelling additional covariate information, the data generation step of Algorithm 1 first requires adequate specification of the associated regression parameters. Since it may be difficult to formulate an 'educated guess' regarding their values, we suggest the use of the process described in Section 4.3. This allows the investigator to choose values of the probabilities of interest such as in Tables 4.3.1 and 4.3.2, and use that information to specify the corresponding the regression parameters, as in Table 4.3.3.

For the odds-ratio, we obtained Monte Carlo estimates of the sample quantities of interest parametrically since we were only considering binary random variables. Simulating Bernoulli trials with known parameters (specified by the investigator) was sufficient to generate binary outcome misclassification. However, in this section, we are allowing for a wide variety of covariates that will have unknown distributions in practice. Hence, we will add a bootstrap step to the data generation algorithm, that will generate the additional covariate data by resampling the original EHR information at each iteration while continuing to conduct Bernoulli trials to generate outcome misclassification. Note that this additional step should not affect convergence behaviour provided that the data are independent and identically distributed and the bootstrap is justifiable for use. The R code for running this algorithm is specified in Appendix C.

With this in mind, we now produce validation sample size determination tables for parameter sets 1 and 3, to mirror the results of Tables 4.3.4 - 4.3.7. This will allow us to observe how well the resulting tables approximate the simulation results. Recall that the purpose of these tables is to allow investigators to select a validation sampling scheme and approach that will approximate the ideal estimator.

Table 4.4.1: Monte Carlo sample size determination for parameter set 1 from Table 4.3.3 with n = 10,000.

Approach	2S	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
c	m	m^*	m^*	m^*	m^*	m^*	m^*	m^*	m^*	m^*
0.2	4056	9158	4925	3870	3380	3202	3266	-	-	-
0.21	3654	8122	4400	3414	2967	2807	2868	3187	-	-
0.22	3301	7237	3943	3029	2623	2480	2536	2830	-	-
0.23	2991	5789	3546	2704	2336	2207	2259	2528	-	-
0.24	2718	5335	3202	2428	2095	1905	2026	2271	2888	-
0.25	2479	4921	2903	2192	1833	1738	1773	2052	2623	-
0.26	2269	4544	2642	1990	1678	1590	1624	1863	2392	-
0.27	2084	4201	2414	1770	1541	1459	1491	1656	2189	-
0.28	1919	3889	2215	1634	1418	1342	1373	1530	2010	-
0.29	1738	3607	2039	1511	1309	1238	1267	1416	1853	-
0.3	1619	3351	1884	1401	1212	1146	1173	1314	1713	-
0.31	1511	3119	1746	1302	1125	1063	1089	1222	1589	-
0.32	1413	2909	1599	1213	1047	989	1014	1138	1479	-
0.33	1323	2718	1497	1132	976	923	946	1063	1380	2376
0.34	1241	2544	1403	1059	913	863	885	995	1277	2235
0.35	1166	2387	1318	993	856	809	830	934	1201	2105
0.36	1097	2243	1239	933	804	760	780	878	1131	1986
0.37	1034	2112	1168	878	757	715	734	827	1067	1875
0.38	977	1992	1102	828	714	675	692	780	1008	1773
0.39	924	1881	1041	782	674	637	654	737	954	1679
0.4	875	1780	986	740	635	601	619	698	904	1593
0.41	830	1687	934	701	603	571	585	662	858	1512
0.42	788	1601	887	666	573	542	556	629	815	1438
0.43	749	1522	843	633	545	516	530	598	775	1369
0.44	713	1438	802	601	519	491	504	568	739	1305
0.45	680	1372	765	573	495	468	481	542	705	1245
0.46	649	1311	729	547	472	447	459	518	673	1190
0.47	620	1253	697	523	451	427	439	495	644	1138
0.48	592	1199	666	500	431	409	420	474	616	1090
0.49	567	1148	638	479	413	391	402	454	590	1044
0.5	544	1100	611	459	396	375	386	436	566	1002
0.51	522	1055	584	440	380	360	370	418	544	962
0.52	501	1013	561	422	365	346	356	402	522	925
0.53	481	973	539	406	350	332	342	386	502	890
0.54	463	936	519	390	337	319	329	372	483	857
0.55	445	900	499	375	324	307	317	358	465	826
0.56	429	867	481	362	312	296	305	345	449	795
0.57	413	835	463	348	301	286	294	333	433	767
0.58	399	805	447	336	290	275	284	321	418	741
0.59	385	777	431	324	280	266	274	310	404	717
0.6	372	750	416	313	270	257	265	300	391	693
0.61	359	724	402	302	261	248	256	290	378	671
0.62	347	700	389	292	253	240	248	281	366	650
0.63	336	677	376	283	245	232	240	272	354	630
0.64	325	655	364	274	237	225	232	263	344	611
0.65	315	634	352	265	229	218	225	255	333	592

Table 4.4.2: Monte Carlo sample size determination for parameter set 3 from Table 4.3.3 with n = 10,000.

Approach	2S	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
c	m	m^*	m^*	m^*	m^*	m^*	m^*	m^*	m^*	m^*
0.2	4122	9474	5095	4001	3439	3189	3180	3442	-	-
0.21	3727	8525	4592	3549	3029	2802	2792	3031	-	-
0.22	3377	7687	4147	3163	2686	2479	2471	2687	-	-
0.23	3068	6950	3754	2833	2397	2210	2202	2397	2965	-
0.24	2795	6302	3408	2551	2153	1911	1906	2153	2671	-
0.25	2554	5733	3103	2308	1878	1745	1740	1879	2418	-
0.26	2341	5233	2834	2099	1723	1598	1593	1724	2199	-
0.27	2152	4793	2597	1917	1585	1467	1462	1585	2009	-
0.28	1985	4404	2387	1715	1461	1350	1345	1461	1842	-
0.29	1836	4061	2202	1590	1350	1246	1242	1350	1660	-
0.3	1703	3608	2037	1477	1251	1154	1149	1250	1543	2601
0.31	1559	3377	1890	1375	1162	1071	1067	1161	1437	2420
0.32	1459	3164	1758	1282	1082	997	993	1081	1341	2257
0.33	1368	2968	1640	1198	1010	930	927	1009	1253	2108
0.34	1285	2788	1534	1122	945	870	867	944	1173	1973
0.35	1208	2622	1422	1053	886	816	813	885	1101	1851
0.36	1138	2470	1340	990	833	767	764	832	1035	1739
0.37	1074	2330	1264	932	784	722	719	783	974	1637
0.38	1015	2201	1194	880	740	681	678	738	919	1544
0.39	960	2082	1130	831	699	643	638	698	869	1458
0.4	910	1972	1070	787	662	607	604	660	822	1379
0.41	863	1870	1015	746	627	576	574	623	779	1307
0.42	820	1776	964	708	594	547	545	592	740	1240
0.43	780	1689	917	674	565	521	518	564	703	1179
0.44	743	1608	873	641	539	496	494	537	669	1115
0.45	709	1533	833	611	514	473	471	512	638	1063
0.46	677	1463	795	582	490	451	449	489	608	1015
0.47	647	1398	759	556	469	431	429	467	580	970
0.48	619	1337	726	533	448	413	411	447	554	928
0.49	593	1280	695	510	429	395	393	428	531	889
0.5	567	1226	667	489	412	379	377	410	509	851
0.51	544	1176	639	469	395	363	362	393	488	816
0.52	523	1125	614	451	379	349	347	377	469	784
0.53	503	1081	590	433	364	335	334	363	450	753
0.54	484	1041	567	417	350	322	321	349	433	723
0.55	466	1002	545	401	337	310	309	336	417	696
0.56	449	966	525	386	325	299	297	323	401	670
0.57	433	931	506	372	313	288	287	312	387	645
0.58	417	898	489	359	302	278	276	300	373	622
0.59	403	867	472	347	292	268	267	290	360	600
0.6	389	838	456	335	282	259	258	280	347	579
0.61	376	809	440	324	272	250	249	270	336	560
0.62	364	783	426	313	263	242	241	262	324	541
0.63	352	757	412	303	255	234	233	253	314	523
0.64	341	733	399	293	247	227	226	245	304	506
0.65	330	710	387	284	239	220	219	237	294	490

Tables 4.4.1 and 4.4.2 demonstrate the approximate relationship between the bound on a Bonferonni interval with $100(1-\alpha)\%$ family-wise confidence level, c, the validation sample size m, and the values of P_1 for the 3S approach. These tables are based on an original EHR data sample size of $n = 10,000$ with 50% in each drug utilization group. Results are presented for values of c ranging from 0.2 to 0.65. In practice, we are likely interested in a particular range of possible validation sample sizes relative to sampling cost constraints, and as such we point out that the range presented here is for demonstrative purposes and can be altered to the desire of the investigator. The minimal required validation sample size for the 2S approach is denoted by m and the 3S approach by m^*.

In Table 4.3.4a we have the following estimates of the standard error of $\hat{\beta}$ for parameter set 1; 0.1 for the 2S approach, 0.1044 for the 3S approach near the incidence rate, and 0.0872 for the 3S approach corresponding with P_1 selected to minimize the estimated variance. We can use these values to determine the half length of the Bonferonni intervals. First, note that $z_{\alpha/2p} = 2.80704$ $(p = 10, \alpha = 0.05)$ which gives us $2.80704 \times 0.1 = 0.2807$ for the 2S approach, $2.80704 \times 0.1044 = 0.2931$ for the 3S approach at the incidence rate and $2.80704 \times 0.0872 = 0.2447$ for the 3S approach at the minimum. Table 4.3.4a uses a validation sample size of $m = 2000$ and referring to Table 4.4.1 under the 2S approach, we appear to require a validation sample size of $m = 2000$ somewhere between values of $c = 0.27$ and 0.28. This corresponds with the simulation based interval half-length of 0.2807.

Next, the incidence rate for parameter set 1 is $\frac{n_1}{n} \approx 0.232$ and we recorded the minimum at $P_1 = 0.4$. Hence, we can again refer to Table 4.4.1 under the columns marked 0.2 and 0.4 to ascertain the values of c corresponding with $m = 2000$. Under the 0.2 column, we see that it occurs between $c = 0.29$ and 0.30 and under the 0.4 column it occurs between $c = 0.24$ and 0.25. Again, the interval half-lengths we computed are 0.2931 for the incidence rate approach and 0.2447 for the minimum approach. Hence, the results produced by the sample size determination algorithm appear to do quite well at approximating these values. Similar results were obtained for the other simulation parameter sets as well as the other values of m.

For Table 4.4.2 we will consider the results for $m = 3,000$ in Table 4.3.7a. The estimates of standard error are 0.0854, 0.0911 and 0.0735 for the 2S, 3S at the incidence rate and 3S at the minimum respectively. These correspond to interval half-lengths of, $2.80704 \times 0.0854 = 0.2397$, $2.80704 \times 0.0911 = 0.2557$ and $2.80704 \times 0.0735 = 0.2063$. Turning to Table 4.4.2, we see $m = 3,000$ occurs between $c = 0.23$ and 0.24 under the 2S column, $c = 0.25$ and 0.26 under the 0.2 column (the incidence rate is $\frac{n_1}{n} = 0.244$) and $c = 0.20$ and 0.21 under the 0.5 column (the minimum). Once again, the results in Table 4.4.2 approximate the interval half-lengths reasonably well. Recall, however, that we restricted the simulation study to ignore those scenarios in which the subsample sizes were overly small. For parameter set 1, we ignored the scenarios for any value of P_1 greater than 0.4. In other words, in our simulation results we do not see the resulting estimates under the 3S approach where $P_1 = 0.5$. In Table 4.4.1, it appears that the ideal value is somewhere close to 0.5 since the corresponding value of c will be between 0.23 and 0.24 (as opposed to 0.24 and 0.25 as previously discussed).

This is an important observation and limitation of these tables. The ideal values of m, c and P_1 may not produce adequate subsample sizes and as such should not be used. To demonstrate this, we will consider parameter set 3 and ignore the effect of Z_2, as it will average out across the groups. We can compute the expected number of observations in each of the binary subgroups given that we know the true parameters/probabilities. In Table 4.4.3 we produce the joint probabilities of the sixteen possible combinations of $(\tilde{a}, a, d, z_1)$.

Table 4.4.3: Expected subgroup sample sizes for parameter set 3 from Table 4.3.3.

$\tilde{A}$	A	D	Z_1	$P(d)$	$P(z_1)$	$\pi_d(z_1)$	$\theta_{da}(z_1)$	$P(\tilde{a}, a, d, z_1)$	$E(n_{\tilde{a},a,d,z_1})$
1	1	1	1	0.5	0.6	0.25	0.9087	0.0682	682
1	1	1	0	0.5	0.4	0.25	0.9	0.045	450
1	1	0	1	0.5	0.6	0.1	0.9271	0.0278	278
1	1	0	0	0.5	0.4	0.1	0.92	0.0184	184
1	0	1	1	0.5	0.6	0.75	0.1094	0.0246	246
1	0	1	0	0.5	0.4	0.75	0.1	0.015	150
1	0	0	1	0.5	0.6	0.9	0.0877	0.0237	237
1	0	0	0	0.5	0.4	0.9	0.08	0.0144	144
0	1	1	1	0.5	0.6	0.25	0.0914	0.0069	69
0	1	1	0	0.5	0.4	0.25	0.1	0.005	50
0	1	0	1	0.5	0.6	0.1	0.073	0.0022	22
0	1	0	0	0.5	0.4	0.1	0.08	0.0016	16
0	0	1	1	0.5	0.6	0.75	0.891	0.2004	2004
0	0	1	0	0.5	0.4	0.75	0.9	0.135	1350
0	0	0	1	0.5	0.6	0.9	0.9123	0.2463	2463
0	0	0	0	0.5	0.4	0.9	0.92	0.1656	1656

The first four columns display all possible events made up by the four binary variables. Next, the column $P(d)$ denotes the probability of the drug utilization group $d = 0, 1$, which is fixed at 50% in each. Recall, that we used a binomial random generator to obtain z_1 with parameter $P(z_1) = 0.6$. For the column marked $\pi_d(z_1)$, these probabilities were all specified in Table 4.3.1, with the exception of the $\pi_1(0)$'s. However these are easily derived since $\pi_1(0) = \frac{e^{\alpha+\psi_1}}{1+e^{\alpha+\psi_1}} = \pi_0(0) = 0.1$ and $\psi_1 = 0$. For the θ-parameters, the values in the $z_1 = 0$ rows are all specified in Table 4.3.2 as well as $\tau_1 = 0.1$. However, to derive the other values, recall that $\theta_{da}(1) = \frac{e^{\kappa_{da}+\tau_1}}{1+e^{\kappa_{da}+\tau_1}}$, where $\kappa_{da} = \log \frac{\theta_{da}(0)}{1-\theta_{da}(0)}$. Finally, the column marked $P(\tilde{a}, a, d, z_1)$ denotes the joint probabilities, or

$$\begin{aligned} P(\tilde{A} = \tilde{a}, A = a, D = d, Z_1 = z_1) &= P(\tilde{A} = \tilde{a} | A = a, D = d, Z_1 = z_1) \\ &\quad \times P(A = a | D = d, Z_1 = z_1) P(D = d) P(Z_1 = z_1), \end{aligned}$$

which is the product of the three middle columns. The last column is the expected number of observations in each category.

Note that the set of misclassified observations in the $\tilde{A} = 0$ group have small expected counts, particularly, $(\tilde{a} = 0, a = 1, d = 0, z_1 = 0)$ with only 16 observations expected in the original data. Also, there appears to be 7,630 expected observations with $\tilde{a} = 0$, so that $P(A = 1, D = 0, Z_1 = 0|\tilde{A} = 0) \approx 0.0021$. Thus, capturing information in our validation sample for these types of scenarios may lead an investigator to choose a slightly lower value of P_1 than would be optimal. To continue with this example, the parameter that will be affected for parameter set 2 is likely κ_{01}, which is the logit of $\theta_{01}(z_1)$. Returning to Table 4.4.3, there are 38 expected observations in the $\tilde{A} = 0$ portion of this group, which translates to a conditional probability of 0.00499. Thus, for an overall validation sample size, $m = 3,000$ and $P_1 = 0.6$, we would expected to draw 5.988 observations in the $(\tilde{a} = 0, a = 1, d = 0)$ validation samples subgroup; for $P_1 = 0.5$ and 0.4, we would expect to draw 7.485 and 8.982 observations respectively. Returning to Table 4.4.2 with a value of $c = 0.2$, the ideal value of P_1 is 0.5 with a validation sample size of 3253; however, since the expected count in this subgroup is 7.485, it may be worth reducing P_1 to 0.4 or 0.3.

We recommend the use of such calculations when utilizing Algorithm 1 for logistic regression with binary outcome misclassification. The calculations were not overly complex given that the scenarios considered were based on relatively simple models. As more covariates are incorporated into the models and more complex modelling structures are implemented, determination of the expected counts may become computationally intensive. We also relied on a bootstrap approach for the data generation step of the algorithm which may be problematic, depending on the complexity of the underlying distribution of the covariates. This could occur, for example, if the values of a particular covariate were to be correlated, perhaps to due time dependency. Nonetheless, there appears to be a wide range of realistic scenarios in which the validation sample size determination algorithm will allow investigators to construct sampling schemes that will balance subsample sizes and variability using the 2S or 3S approaches.

4.5 Conclusions and Discussion

This discussion has focused on extending the work in Chapters 2 and 3 to allow for the inclusion of multiple covariates when modelling the relationship between drug utilization and adverse events in EHRs with the presence of outcome misclassification. We have shown that the 2S likelihood presented in Lyles et al. (2011) [25] will produce MCB-adjusted estimates of the logistic regression parameters of interest using a random sampling approach to draw the validation data. We introduced the 3S likelihood approach and demonstrated the use of conditional sampling to adjust for outcome misclassification, while allowing for more control by the investigator over the categorical structure of the validation sample. We demonstrated through Monte Carlo simulation that the potential for complexity in modelling the π and θ-parameters makes the ARE of these estimators difficult to predict when planning a validation sampling scheme. Thus, modelling these approaches under a multi-sample framework allowed for the implementation of a similar validation sample size determination algorithm as in Section 3, with a few adjustments.

The first notable adjustment is that we suggest bootstrapping the original covariate information for the data generation step of the algorithm. This may cause problems in scenarios in which the bootstrap approach does not perform well. As such future work could consider methods to account for these types of situations. For example, there may be dependencies in the data due to some unobserved attribute, such as clustering of results by health care institution. The bootstrap may fail to approximate the limiting distribution of the covariate data in this scenario. Our second recommendation is to attempt to ascertain the expected counts in the smaller misclassification groups of the validation sample based on the specified target parameter values. This will assist in selecting realistic values of P_1 that, coupled with the sample size determination tables, will return sufficient data in all subgroups. However, we have not addressed the possible impact of error in the implementation of this procedure. If the parameters are misspecified, the validation sample size determination table will produce poor approximations of the relationships between m, P_1 and c. The extent to which this will affect the estimators is more difficult to predict than in the contingency table methods. We saw similarity in the relationships between the 2S and 3S approaches across most sets of simulation parameters in the results in Chapter 3; however, this is not the case for the results in this section. As such, future work is needed to assess the impact of specifying incorrect parameters on the resulting numerical approximations.

Finally, we have outlined two approaches for using internal validation data to address the problems of binary outcome misclassification. This work was motivated by EHR data, as the AE occurrences in these records can be dichotomized in the observational interval, allowing for application of binary regression. However, we are also interested in exploring the temporal aspects of such data and this problem is addressed in the next chapter.

Chapter 5

A Validation Sampling Approach for Unbiased Estimation in Misclassified Right Censored Continuous Time Survival Data With and Without Competing Risks

5.1 Introduction

Electronic health record (EHR) datasets are comprised of person-oriented longitudinal data on health service utilization and health outcomes, and, as such has an inherent time-to-event structure. In this section we consider a time-to-event framework to model EHR data with the presence of outcome misclassification.

To begin the discussion, we must reconsider the manner in which outcome misclassification appears in the data. In Chapter 1, we noted a number of ways that these errors can occur, such as coding errors, variation in AE definitions, imperfect diagnostic tests as well as misclassification due to the difference in time between AE occurrence and the time to possible entry in the EHR datasets. Within a time-to-event framework, the latter type of error presents a particular modelling challenge that requires careful consideration of how right censorship appears in the data. Each patient's EHR contains a large amount of information regarding interactions with the health system. As such censored observations may be present as an encounter that takes place after the end of the observational interval, representing a loss to follow up (LTF) type event. Such events may be non-diagnostic encounters such as lab tests or the filling of a prescription. Alternatively, the LTF event may be an interaction with a physician or other health professional in which the patient is deemed healthy or has a competing AE after which the observational interval ends, perhaps due to the patient switching his or her health service provider.

Misclassification in either of these scenarios would present as the incorrect observation of the AE of interest occurring prior to censorship in the EHR. Thus, we would observe an event in the data that results in the recording of an incorrect diagnosis. However, the diagnosis of the AE of interest in the first type of setting (conducting a lab test, for example) is unlikely to occur since there is no diagnostic side to the encounter. Considering the second scenario, the AEs of interest in pharmacovigilance are generally serious in nature. Thus, mistakingly diagnosing an individual as healthy when they are in fact experiencing a serious event is unlikely. Alternatively, the incorrect diagnosis of an AE other than that of interest would appear to be a plausible form of outcome misclassification. However, it is better modelled as the occurrence of a competing risk than a censorship event. Hence, we will make the simplifying assumption that misclassification is only possible if a 'true' AE has occurred, ignoring the alternate AE diagnosis at first, and introduce a competing risks framework to model the other possibility later in Section 5.2.2.

The simplifying assumption restricts this discussion to only certain observed events being classified erroneously. If we are to observe an AE occurrence in the original EHR data, we can assume it to be true, however, the cause-specific AE type may be incorrect. Alternatively, an observed censored observation may have been incorrectly censored and we failed to observe an AE occurrence that actually occurred. In other words, the only possible observed temporal errors will manifest as censored events in the data since misclassification is only possible if a patient should have been truthfully classified as having had an AE occurrence.

The primary interest in pharmacovigilance is to model the risk of an adverse health outcome as a function of the utilization of the drug of interest. Thus, we will model the hazard as a function of this binary covariate as well as a function of other covariate information. However, since we are modelling the time to first AE occurrence in continuous time, we will restrict the additional explanatory information and misclassification rates to being independent of time. Allowing for time-dependent covariates adds a large complexity to the discussion as the AE occurrence rates depend on this information while simultaneously being influenced by misclassification, which may also be modelled as a function of time. Relaxing these assumptions will require careful specification of these relationships and will not necessarily allow for general modelling, hence we do not consider this here. However, this assumption is nonetheless restrictive and needs to be investigated in future work.

We begin by introducing a univariate parametric survival model in which we assume a constant baseline hazard and outcome misclassification rate. In this model, misclassification will appear in the data as incorrectly observed censorship type events. To extend this basic model further, we will allow for the presence of competing AEs, since in the first model, these events may be considered as censored observations. We employ a cause-specific (C-S) hazards approach to model the competing events in a parametric setting (Kalbfleisch and Prentice

(2002) [19]). Thus, misclassification can occur either by erroneously observing the competing AE type, or by misclassification due to a delayed diagnosis. Thus, we propose a second model that allows for both misclassification of event time as well as competing AE type.

This chapter is organized as follows. The theoretical development needed to outline the previously-discussed models is presented in Section 5.2. The asymptotic properties of the resulting estimators using the multi-sample framework described in Appendix A. In Section 5.3 we model the time to AE occurrence under an exponential distribution with hazard function depending on drug utilization status as well as an additional binary covariate and positive continuous covariate. In Sections 5.6-5.8, we will conduct Monte Carlo simulations for both models to numerically investigate the properties of the resulting estimators. We will also consider the effect of validation sample size and its relationship with the original EHR data sample size on the bias adjusted estimators under these models. We discuss the practical implications of this work and discuss possible future contributions in Section 5.9.

5.2 Notation and Likelihood Construction

Let T be a random variable denoting the time to first AE occurrence in the observational interval, $[0, \tau]$, with density $f(t)$ and survival function, $S(t)$, and let C denote the censorship time. Under perfect classification, we observe $X = min(T, C)$ and can describe the event using the indicator $\delta = \mathbf{1}_{\{T<C\}}$. Thus, the hazard function at time t is,

$$\lambda(t; d, z) = \lim_{dt \to 0} \frac{P[T \in I_{dt} | T \geq t, D = d, Z = z]}{dt},$$

where I_{dt} denotes the infinitesimal interval $[t, t + dt)$, D is an indicator of drug utilization status, where $D = 1$ if the patient under study is taking the drug of interest and 0 otherwise, and Z denotes a vector of the additional covariate information for the patient under study that is independent of t. For models incorporating the presence of competing AEs, we will denote the j^{th} cause-specific hazard as,

$$\lambda_j(t; d, z) = \lim_{dt \to 0} \frac{P[T \in I_{dt}, J = j | T \geq t, D = d, Z = z]}{dt}.$$

For the remainder of this discussion, if competing AEs are considered, we will consider the possibility of two competing risks. The first ($J = 1$) is the AE of interest and the second ($J = 2$) will represent all other AEs.

As we are considering parametric models the baseline hazards are assumed to be constant. Finally, to introduce misclassification, we will use a 'tilde' to denote the observed, possibly misclassified, versions of the previously introduced random variables. For instance, $\tilde{T}$ denotes the time to first AE occurrence observed in the EHR, while T denotes the unobserved 'true' time to adverse event occurrence. We can then introduce the following probabilities,

$$\begin{aligned} \theta(t) &= P[\tilde{T} \in I_{dt} | T \in I_{dt}, D = d, Z = z], \\ \xi_j(t) &= P[\tilde{J} = j | J = j, T \in I_{dt}, D = d, Z = z] \end{aligned}$$

and

$$\phi_j(t) = P[\tilde{T} \in I_{dt} | T \in I_{dt}, \tilde{J} = j, D = d, Z = z],$$

for $j = 1, 2$. Recall, that we assume that these probabilities are time-homogeneous and to further simplify, we will have them not depend on D and Z. To model these probabilities as functions of the covariates D and Z, we could use a logistic transformation to incorporate the information into the likelihoods described in the following sections, however, for simplicity we will denote them as $\theta(t) = \theta$,

$\xi_j(t) = \xi_j$ and $\phi_j(t) = \phi_j$. Further, it is reasonable to assume that $\phi_1 = \phi_2$ since the observed C-S event type should not effect the probability that the time to event is correctly recorded in the EHR; hence, $\phi = \phi_1 = \phi_2$.

With this notation in mind, we will outline each of the models described in Section 5.1. For clarity, we will denote the model number using a superscript in brackets and will refer to the individual sample (original or validation) using a subscript. For example, the validation sample likelihood for model 2 will be denoted, $L_V^{(2)}$. Next, recall the assumption that a 'true' AE occurrence must take place in order for misclassification to be possible will hold throughout the remainder of the chapter; the probability can be written as, $P[\tilde{T} > t | T > t] = 1$ for all t. Finally, for brevity, we present the theoretical results in this chapter without justification; proofs are outlined in Appendix B.

5.3 A Univariate Model, $L^{(1)}$

Since we are only considering the AE of interest as a failure, we can describe the original EHR data likelihood, $L_O^{(1)}$, by the following probabilities,

$$P[\tilde{T} \in I_{dt} | D = d, Z = z] \quad = \quad \theta f(t; d, z) dt \tag{5.1}$$

and

$$P[\tilde{T} > t | D = d, Z = z] \quad = \quad 1 - \theta(1 - S(t; d, z)).$$

These lead to the following observed data likelihood,

$$L_O^{(1)} \propto \prod_{i=1}^{n} [\theta f(t_i; d_i, z_i)]^{\tilde{\delta}_i} [1 - \theta(1 - S(t_i; d_i, z_i))]^{1 - \tilde{\delta}_i}.$$

According to our misclassification assumption, errors can only occur if a true AE occurrence has taken place. This implies that we are only able to observe error in the subset of the data that was originally observed to be censorship events, $V_0 = \{i : \tilde{\delta}_i = 0\}$, $i = 1, ..., n$. Thus, we will draw a validation sample of size m from V_0. The likelihood, $L^{(1)}_{V_0}$ for this subsample will be composed of the following probabilities,

$$P[T \leq t | \tilde{T} > t, D = d, Z = z] \quad = \quad \frac{(1-\theta)(1 - S(t; d, z))}{1 - \theta(1 - S(t; d, z))} \tag{5.2}$$

and

$$P[T > t | \tilde{T} > t, D = d, Z = z] \quad = \quad \frac{S(t; d, z)}{1 - \theta(1 - S(t; d, z))}.$$

The validation sample likelihood is,

$$L^{(1)}_V \propto \prod_{\{v: v \in V_0\}} \left[\frac{(1-\theta)(1 - S(t_v; d_v, z_v))}{1 - \theta(1 - S(t_v; d_v, z_v))}\right]^{\delta_v(1-\tilde{\delta}_v)} \left[\frac{S(t_v; d_v, z_v)}{1 - \theta(1 - S(t_v; d_v, z_v))}\right]^{(1-\delta_v)(1-\tilde{\delta}_v)} \tag{5.3}$$

giving us the complete likelihood as $L^{(1)} = L^{(1)}_O \times L^{(1)}_V$, so that

$$\begin{aligned} L^{(1)} \quad & \propto \quad \prod_{i=1}^{n} [\theta f(t_i; d_i, z_i)]^{\tilde{\delta}_i} [1 - \theta(1 - S(t_i; d_i, z_i))]^{1-\tilde{\delta}_i} \qquad (5.4) \\ & \quad \times \prod_{\{v: v \in V_0\}} \left[\frac{(1-\theta)(1 - S(t_v; d_v, z_v))}{1 - \theta(1 - S(t_v; d_v, z_v))}\right]^{\delta_v(1-\tilde{\delta}_v)} \left[\frac{S(t_v; d_v, z_v)}{1 - \theta(1 - S(t_v; d_v, z_v))}\right]^{(1-\delta_v)(1-\tilde{\delta}_v)}. \end{aligned}$$

Note that we are treating δ in $L^{(1)}$ as though it is a misclassification indicator. This follows from the assumption that the validation sample information is assumed to be based on some 'gold standard', or an infallible diagnostic test, which implies that it produces the same indication when paired with $\tilde{\delta}$.

5.4 A Competing Risks Model, $L^{(2)}$

For this model, we will introduce the possibility of competing AEs; however, with the exception of the AE of interest, all others are grouped into an 'other' category leaving two competing events, the AE of interest and all other competing AEs combined. Thus, we will have two possible forms of misclassification. The first is similar to that discussed in Section 5.2.1, where an individual is observed as being incorrectly censored. For instance, a patient visit to his/her doctor to discuss test results may result in the conclusion that further testing is required prior to finalizing a diagnosis, followed by the patient seeking a second opinion elsewhere. However, the patient was actually experiencing the AE of interest during this visit. In addition, we are allowing for the possibility that the conclusion of the encounter is an incorrect diagnosis of a competing AE. This additional type of misclassification can be thought of as correctly concluding that an AE occurred; however, the C-S AE type was erroneously observed. This type of diagnostic error is likely to appear in EHRs as people present themselves at a hospital for the occurrence of groups of symptoms. For instance, since EHRs include emergency room visits and in the limited diagnostic time frame associated with these extreme encounters, misdiagnoses may be more probable. Hence, misdiagnoses are also possible in both $\tilde{\delta}$ groups unlike the model discussed in Section 5.2.1. The only restriction to misclassification is the assumption that underlies all the models, which is that an AE of any type must have truly occurred for possible misclassification resulting in the observation of either a censorship event or the incorrect AE to be present in the EHRs.

To construct the likelihood for the above, we must consider the contribution of the observed competing event information in the original EHR data. Since we can reasonably assert that the presence of an observed competing event type will not effect the probability of misclassification and we have modelled the ϕ and ξ-parameters to be independent of the additional covariate information, we can write $P[\tilde{T} \in I_{dt}, \tilde{J} = \tilde{j} | D = d, Z = z] = \phi f(t; d, z) P[\tilde{J} = j | T \in I_{dt}, D = d, Z = z]$, for $\tilde{j} = 1, 2$, which allows us to specify the probabilities making up the original likelihood, $L_O^{(2)}$, as,

$$\begin{aligned} P[\tilde{T} \in I_{dt}, \tilde{J} = 1 | D = d, Z = z] &= \phi S(t; d, z)[\xi_1 \lambda_1(t; d, z) + (1 - \xi_2)\lambda_2(t; d, z)]dt, \quad (5.5) \\ P[\tilde{T} \in I_{dt}, \tilde{J} = 2 | D = d, Z = z] &= \phi S(t; d, z)[(1 - \xi_1)\lambda_1(t; d, z) + \xi_2 \lambda_2(t; d, z)]dt \end{aligned}$$

and

$$P[\tilde{T} > t|D = d, Z = z] \quad = \quad 1 - \phi(1 - S(t; d, z)).$$

The original data likelihood is then,

$$\begin{aligned} L_O^{(2)} \;\propto\; & \prod_{i=1}^{n} [\phi S(t_i; d_i, z_i)[\xi_1 \lambda_1(t_i; d_i, z_i) + (1 - \xi_2)\lambda_2(t_i; d_i, z_i)]]^{\tilde{\delta}_i \mathbf{1}_{\tilde{J}_i = 1}} \qquad (5.6) \\ & \times [\phi S(t_i; d_i, z_i)[(1 - \xi_1)\lambda_1(t_i; d_i, z_i) + \xi_2 \lambda_2(t_i; d_i, z_i)]]^{\tilde{\delta}_i \mathbf{1}_{\tilde{J}_i = 2}} [1 - \phi(1 - S(t_i; d_i, z_i))]^{(1 - \tilde{\delta}_i)}. \end{aligned}$$

The validation sample portion of the likelihood is made up of two differing subsamples, each housing information on a different type of misclassification. The first is drawn from $V_1 = \{i : \tilde{\delta}_i = 1\}$ and contains the information on misclassification of competing AE type. Since we assume that T and δ are measured correctly in V_1, at each time t the associated underlying probabilities for an observation selected from V_1 will be,

$$\begin{aligned} P[J = 1|\tilde{J} = 1, T \in I_{dt}, D = d, Z = z] &= \frac{\xi_1 \lambda_1(t; d, z)}{\xi_1 \lambda_1(t; d, z) + (1 - \xi_2)\lambda_2(t; d, z)}, \quad (5.7) \\ P[J = 2|\tilde{J} = 1, T \in I_{dt}, D = d, Z = z] &= \frac{(1 - \xi_2)\lambda_2(t; d, z)}{\xi_1 \lambda_1(t; d, z) + (1 - \xi_2)\lambda_2(t; d, z)}, \\ P[J = 1|\tilde{J} = 2, T \in I_{dt}, D = d, Z = z] &= \frac{(1 - \xi_1)\lambda_1(t; d, z)}{(1 - \xi_1)\lambda_1(t; d, z) + \xi_2 \lambda_2(t; d, z)}, \end{aligned}$$

and

$$P[J = 2|\tilde{J} = 2, T \in I_{dt}, D = d, Z = z] = \frac{\xi_2 \lambda_2(t; d, z)}{(1 - \xi_1)\lambda_1(t; d, z) + \xi_2 \lambda_2(t; d, z)}.$$

To account for misclassification of the true time to AE occurrence, a sample drawn from V_0 as in Section 5.2.1 is required. However, we must extend the validation sample portion of the likelihood described in Section 5.2.1 to account for the additional competing AE event type. In other words, a misclassified censorship event may truly be an AE occurrence, but of either C-S event type. Hence, the following expressions can be specified to model the underlying probabilities,

$$P[T \leq t, J = j|\tilde{T} > t, D = d, Z = z] \;=\; \frac{\lambda_j(t; d, z)}{\sum_{j=1,2}\lambda_j(t; d, z)}\frac{(1-\phi)(1-S(t; d, z))}{1-\phi(1-S(t; d, z))} \tag{5.8}$$

and

$$P[T > t|\tilde{T} > t, D = d, Z = z] \;=\; \frac{S(t; d, z)}{1-\phi(1-S(t; d, z))},$$

for $j = 1, 2$.

The combined likelihood can then be written as,

$$\begin{aligned} L^{(2)} \;\propto\; & \prod_{i=1}^{n}[\phi S(t_i; d_i, z_i)[\xi_1\lambda_1(t_i; d_i, z_i) + (1-\xi_2)\lambda_2(t_i; d_i, z_i)]]^{\tilde{\delta}_i \mathbf{1}_{\tilde{J}_i=1}} \\ & \times[\phi S(t_i; d_i, z_i)[(1-\xi_1)\lambda_1(t_i; d_i, z_i) + \xi_2\lambda_2(t_i; d_i, z_i)]]^{\tilde{\delta}_i \mathbf{1}_{\tilde{J}_i=2}}[1-\phi(1-S(t_i; d_i, z_i))]^{(1-\tilde{\delta}_i)} \\ & \times \prod_{\{v: v\in V_0\}}\left[\frac{\lambda_1(t_v; d_v, z_v)}{\sum_{j=1,2}\lambda_j(t_v; d_v, z_v)}\frac{(1-\phi)(1-S(t_v; d_v, z_v))}{1-\phi(1-S(t_v; d_v, z_v))}\right]^{\delta_v(1-\tilde{\delta}_v)\mathbf{1}_{J_v=1}} \\ & \times\left[\frac{\lambda_2(t_v; d_v, z_v)}{\sum_{j=1,2}\lambda_j(t_v; d_v, z_v)}\frac{(1-\phi)(1-S(t_v; d_v, z_v))}{1-\phi(1-S(t_v; d_v, z_v))}\right]^{\delta_v(1-\tilde{\delta}_v)\mathbf{1}_{J_v=2}} \\ & \times\left[\frac{S(t_v; d_v, z_v)}{1-\phi(1-S(t_v; d_v, z_v))}\right]^{(1-\delta_v)(1-\tilde{\delta}_v)} \\ & \times \prod_{\{v: v\in V_1\}}\left[\frac{\xi_1\lambda_1(t_v; d_v, z_v)}{\xi_1\lambda_1(t_v; d_v, z_v) + (1-\xi_2)\lambda_2(t_v; d_v, z_v)}\right]^{\delta_v\tilde{\delta}_v\mathbf{1}_{(J_v=1,\tilde{J}_v=1)}} \\ & \times\left[\frac{(1-\xi_2)\lambda_2(t_v; d_v, z_v)}{\xi_1\lambda_1(t_v; d_v, z_v) + (1-\xi_2)\lambda_2(t_v; d_v, z_v)}\right]^{\delta_v\tilde{\delta}_v\mathbf{1}_{(J_v=2,\tilde{J}_v=1)}} \\ & \times\left[\frac{(1-\xi_1)\lambda_1(t_v; d_v, z_v)}{(1-\xi_1)\lambda_1(t_v; d_v, z_v) + \xi_2\lambda_2(t_v; d_v, z_v)}\right]^{\delta_v\tilde{\delta}_v\mathbf{1}_{(J_v=1,\tilde{J}_v=2)}} \\ & \times\left[\frac{\xi_2\lambda_2(t_v; d_v, z_v)}{(1-\xi_1)\lambda_1(t_v; d_v, z_v) + \xi_2\lambda_2(t_v; d_v, z_v)}\right]^{\delta_v\tilde{\delta}_v\mathbf{1}_{(J_v=2,\tilde{J}_v=2)}}. \end{aligned} \tag{5.9}$$

This likelihood is the most general form in that it accounts for all allowable types of misclassification discussed thus far. Note that the work of Rompaye et al. (2010) [36] allowed only for misclassification of competing AE type (misclassified cause of death). In addition, they assumed that the observation of the time to event was measured correctly. Rompaye et al. (2010) [36] further assumed that the misclassification probabilities were known, and that the baseline C-S hazards were proportional, thereby allowing for a semi-parametric approach. Under their assumptions, a parametric likelihood would be a simplification of the original data likelihood, $L_O^{(2)}$. As such, we are able to reduce this likelihood to extend Rompaye et al.'s (2010) [36] work to incorporate validation data to produce MCB-adjusted estimation under unknown misclassification rates by first modelling the original data probabilities as,

$$\begin{aligned} P[T \in I_{dt}, \tilde{J} = 1 | D = d, Z = z] &= S(t; d, z)[\xi_1 \lambda_1(t; d, z) + (1 - \xi_2)\lambda_2(t; d, z)]dt, \quad (5.10) \\ P[T \in I_{dt}, \tilde{J} = 2 | D = d, Z = z] &= S(t; d, z)[(1 - \xi_1)\lambda_1(t; d, z) + \xi_2 \lambda_2(t; d, z)]dt \end{aligned}$$

and

$$P(T > t | D = d, Z = z) \;\; = \;\; S(t; d, z).$$

Note, that these are similar to (5.5); however, the ϕ-parameters are not needed. The validation data portion will be (5.7), resulting in the combined likelihood,

$$\begin{aligned} L^{(2b)} \propto & \prod_{i=1}^{n} [S(t; d, z)[\xi_1 \lambda_1(t_i; d_i, z_i) + (1 - \xi_2)\lambda_2(t_i; d_i, z_i)]]^{\delta_i \mathbf{1}_{\tilde{J}_i = 1}} \quad (5.11) \\ & \times [S(t_i; d_i, z_i)[(1 - \xi_1)\lambda_1(t_i; d_i, z_i) + \xi_2 \lambda_2(t_i; d_i, z_i)]]^{\delta_i \mathbf{1}_{\tilde{J}_i = 2}} [S(t_i; d_i, z_i)]^{(1 - \delta_i)} \\ & \times \prod_{\{v : v \in \mathcal{V}_1\}} \left[\frac{\xi_1 \lambda_1(t_v; d_v, z_v)}{\xi_1 \lambda_1(t_v; d_v, z_v) + (1 - \xi_2)\lambda_2(t_v; d_v, z_v)} \right]^{\delta_v \mathbf{1}_{(J_v = 1, \tilde{J}_v = 1)}} \\ & \times \left[\frac{(1 - \xi_2)\lambda_2(t_v; d_v, z_v)}{\xi_1 \lambda_1(t_v; d_v, z_v) + (1 - \xi_2)\lambda_2(t_v; d_v, z_v)} \right]^{\delta_v \mathbf{1}_{(J_v = 2, \tilde{J}_v = 1)}} \\ & \times \left[\frac{(1 - \xi_1)\lambda_1(t_v; d_v, z_v)}{(1 - \xi_1)\lambda_1(t_v; d_v, z_v) + \xi_2 \lambda_2(t_v; d_v, z_v)} \right]^{\delta_v \mathbf{1}_{(J_v = 1, \tilde{J}_v = 2)}} \\ & \times \left[\frac{\xi_2 \lambda_2(t_v; d_v, z_v)}{(1 - \xi_1)\lambda_1(t_v; d_v, z_v) + \xi_2 \lambda_2(t_v; d_v, z_v)} \right]^{\delta_v \mathbf{1}_{(J_v = 2, \tilde{J}_v = 2)}}. \end{aligned}$$

Since $L^{(2b)}$ is clearly a simplification of $L^{(2)}$, we will not investigate this likelihood in the simulation studies described in Section 5.4, however, we do nevertheless choose to present it here as it is as a parametric extension of the likelihood considered by Rompaye et al. (2010) [36].

Deriving the MLEs requires numerical maximization of the corresponding log likelihoods, as they are far too complex to admit closed-form solutions. In later sections, we will use the R function *nlminb*() for this purpose. However, first we will outline the asymptotic properties of these multi-sample likelihoods. To begin, we will demonstrate that both representations of misclassification in EHR data satisfy the definition of multi-sample data as described in Appendix A, Definition A.2.1. We then briefly discuss the asymptotic properties developed in Appendix A and demonstrate that under certain regularity conditions that are likely to be satisfied by EHR data, the inverse of the observed information matrix is a consistent estimator of the asymptotic variance-covariance matrix.

5.5 Asymptotic Properties for Multi-Sample Maximum Likelihood Estimation

In Section 2.3, we demonstrated that modelling the outcome of interest in EHR data as binary under a fixed time assumption satisfies the definition of multi-sample data. Throughout this discussion we have relaxed this assumption and modelled EHR data possessing a time-to-event data structure. Thus, we will verify that these modelling assumptions satisfy Definition A.2.1 in Appendix A allowing us to utilize a multi-sample framework to outline the asymptotic properties of the MLEs derived from $L^{(1)}$ and $L^{(2)}$. For referencing purposes, we will denote these p-dimensional MLEs as, $\hat{\gamma}_1$ and $\hat{\gamma}_2$ respectively.

Consider drawing N patients from the source population of interest and monitor the way these patients interact with the health care institutions under study over time. For the single AE case, $L^{(1)}$, we have observations of the form $(\tilde{x}_i, \tilde{\delta}_i, \delta_i, d_i, z_i)$ for $i = 1, ..., m$ and $(\tilde{x}_i, \tilde{\delta}_i, d_i, z_i)$ for $i = m + 1, ..., N$ where $N = n + m$ and $\tilde{x}$

is the realized value of $\tilde{X}$, the observed version the true value of X. Thus, we are independently drawing two samples from the source population with weights $\left(\frac{m}{N}, \frac{n}{N}\right)$. For the competing risks model, $L^{(2)}$, we are drawing observations with the addition of competing event types. Further, we are selecting three samples, in which two are drawn based on $\tilde{\delta}$-status at observed event time $\tilde{x}$. Thus, our first sample has observations of the form, $(\tilde{x}_i, \tilde{\delta}_i = 1, \delta_i, \tilde{j}, j, d_i, z_i)$ for $i = 1, ..., m_1$, the second $(\tilde{x}_i, \tilde{\delta}_i = 0, \delta_i, \tilde{j}, j, d_i, z_i)$ for $i = m_1 + 1, ..., m_1 + m_0$ and the third $(\tilde{x}_i, \tilde{\delta}_i, \tilde{j}, d_i, z_i)$ for $i = m_1 + m_0 + 1, ..., N$ where $N = n + m_1 + m_0$. In the same way as for the single AE case, the sample weights are $\left(\frac{m_1}{N}, \frac{m_0}{N}, \frac{n}{N}\right)$.

The likelihoods presented in Sections 5.2.1 and 5.2.2 are written in their general forms with respect to functions that describe the distribution of T, namely, the density, survival and hazard functions. However, the distribution of T has not yet been specified and as such, provided that the choice is a commonly used parametric survival model (Kalbfleisch and Prentice (2002) [19]), the regularity conditions listed in Theorem A.2.3 will be satisfied. Next, the validation sample data is gathered by randomly selecting observations from the original EHR data based on their original event status. Then, we verify the encounter and determine whether the observation was misclassified or not. Thus, these portions of the likelihood can be modelled under binomial and multinomial distributions for $L^{(1)}$ and $L^{(2)}$ respectively.

Therefore, our data modelled under both likelihoods satisfy Definition A.2.1, which allows us to apply the results of Theorem A.2.3. Thus, we can model the approach in Section 5.2.1 under a two sample approach and the approach in Section 5.2.2 as a three sample approach. With that in mind, we can turn to many of the results outlined in Appendix A and demonstrate that we have a consistent estimator of the asymptotic variance-covariance matrix via the inverse of the observed information matrix. This is first due to the fact that inference in parametric survival data is done using the score function, which is central to the application of Theorem A.2.3. Specifically, since we have noted that the regularity conditions are satisfied, we have consistency of the estimated parameter vector, $\hat{\gamma} \xrightarrow{p} \gamma_0$ and

$$\sqrt{N}(\hat{\gamma}_N - \gamma_0) \xrightarrow{D} N\left(\mathbf{0}, \left[\sum_{s=1}^{s} \omega_s \mathcal{I}_s(\gamma_0)\right]^{-1}\right), \tag{5.12}$$

where ω_s denotes the s^{th} weight probability in the definition of multi-sample data and $\mathcal{I}_s(\gamma_0)$ is the negative expectation of the Hessian matrix for the s^{th} sample at the hypothesized value, γ_0.

As previously mentioned, as long as we have convergence of the sample weights to probability weights, we can estimate the inverse of the asymptotic variance with,

$$\frac{1}{N}\sum_{s=1}^{S} n_s \bar{I}_s(\hat{\gamma}) \xrightarrow{p} \sum_{s=1}^{S} \omega_s \mathcal{I}_s(\gamma_0), \tag{5.13}$$

where $\bar{I}_s(\hat{\gamma}) = \frac{1}{n_s}\sum_{i=1}^{n_s} I_{si}(\hat{\gamma}) = \frac{1}{n_s}\sum_{i=1}^{n_s}\left[-\frac{\partial l_s(\gamma|x_{si})}{\partial\gamma\partial\gamma^T}\right]_{\gamma=\hat{\gamma}}$ and $N = \sum_{s=1}^{S} n_s$. This can be written as sum of the negative of the Hessian matrices derived from each of the S samples, which verifies the assertion that the inverse of the observed information matrix is a consistent estimator. Thus, we can write,

$$\begin{aligned} \hat{\sigma}^2(\hat{\gamma}_1) &= [n\bar{I}_O(\hat{\gamma}_1) + m\bar{I}_V(\hat{\gamma}_1)]^{-1} \\ &= [I_O(\hat{\gamma}_1) + I_V(\hat{\gamma}_1)]^{-1}, \end{aligned} \tag{5.14}$$

and

$$\begin{aligned} \hat{\sigma}^2(\hat{\gamma}_2) &= [n\bar{I}_O(\hat{\gamma}_2) + m_1\bar{I}_{V_1}(\hat{\gamma}_2) + m_0\bar{I}_{V_0}(\hat{\gamma}_2)]^{-1} \\ &= [I_O(\hat{\gamma}_2) + I_{V_1}(\hat{\gamma}_2) + I_{V_0}(\hat{\gamma}_2)]^{-1}, \end{aligned} \tag{5.15}$$

where $I_s(\hat{\gamma})$ denotes the matrix of negative second derivatives for the s^{th} sample at the MLE, $\hat{\gamma}$.

5.6 Simulation Studies to Investigate the Properties of $L^{(1)}$ and $L^{(2)}$

In this section, we will numerically investigate the asymptotic properties of the MLEs derived in Sections 5.2.1 and 5.2.2 via Monte Carlo simulation. We will begin by specifying distributional assumptions surrounding the time to first AE occurrence, T. As our goal is to realistically model EHR data; hence we will consider an exponential distribution with rate,

$$\lambda(t; d, z_i) = \lambda_0 \exp(\beta d + \psi^T z_i),$$

and

$$\begin{aligned}\lambda_1(t; d, z_i) &= \lambda_{01} \exp(\beta_1 d + \psi^T z_i) \\ \lambda_2(t; d, z_i) &= \lambda_{02} \exp(\beta_2 d + \eta^T z_i),\end{aligned}$$

to describe the C-S hazards. In other words, ψ will denote a vector of the regression coefficients for the hazard of the AE of interest at time t regardless of the presence of competing events or not. Thus, we can write $L^{(1)}$ and $L^{(2)}$ with respect to these distributional assumptions and solve the MLEs by maximizing the log-likelihoods.

For our simulation studies, we will use these models with two additional covariates beyond drug utilization status. The first, Z_1, will be a binary random variable, such as gender, while the second, Z_2, will be a positive continuous random variable that could denote a key biochemical marker related to a health outcome. Recall, that we are not considering time dependent covariates in this discussion.

Considerations regarding the selection of simulation parameters differ for the two methods. In Section 5.2.1, we are considering a two sample approach as we are only drawing validation information from V_0. Hence, we must be able to observe an adequately large sample size for this set. The reality of EHR data is such that this is an easy data requirement to satisfy, since most AEs will be rare relative to the overall dataset. For the competing risks approach, we require samples drawn from both V_1 and V_0, which implies that we need a reasonably large number of observations drawn from the V_1 subsample. As such, we must take care in selecting parameter sets for the simulations scenarios.

To investigate the properties of $\hat{\gamma}_1$, we will use standard data generation algorithms to generate right censored survival data, which entails the generation of both observations under the distribution of T, as well as under the censorship distribution, C, followed by selection of the observation with smaller event time. Hence, we will select another exponential distribution with rate ν for C and can alter the value of ν in order to assert control over the rates of failures and censorships. Next, we will select misclassification rates that are low and allow them to increase, so as to observe the effect on parameter estimation. Note, that this data generation process creates latent event times with δ indicating which type of event has occurred (AE or censorship) as the smaller event time. However, misclassification is implemented by erroneously returning the latent (later) event time, given that the correct (earlier) event time is associated with a true AE occurrence, where these observations are selected via a Bernoulli trial with the misclassification rate as its parameter. Thus, misclassification is generated in the same way that it would appear in EHR data in that we are observing a relationship between two event times in which we fail to observe the event time of interest (which would end the study interval) and incorrectly observe the later loss to follow-up time.

For $\hat{\gamma}_2$, we begin by applying the data generation algorithm suggested by Beyersmann et al. (2009) [3], in which generation begins by specification of the C-S hazards. Generation of T and C occur in the same manner as before, although the all-cause hazard is used to generate T. Next, the C-S types are applied to each observation in which $\delta = 1$ by conducting a binomial experiment where the success probability is the proportion that the corresponding C-S hazards contribute to the overall all-cause hazard. Now, consider that misclassification of competing event types brings an additional form of classification error. This expresses itself in the validation sample portion of the likelihood, $L_{V_1}^{(2)}$, as being similar to binary misclassification error, in that we are considering the incorrect C-S AE type being applied to the AE occurrence at a particular time. Recall that our initial assumption regarding misclassification will imply that an event observed at time t as an AE will be observed without error, which justifies this assertion. Since time does not effect this part of the likelihood as opposed to the validation sample drawn from V_0, we may hypothesize that smaller validation sample sizes are needed from V_1. We will investigate this as well by altering the sizes of the validation samples drawn from V_1, m_1, and V_0, m_0.

In Tables 5.7.1 and 5.8.1 in the following sections, we list all the simulation parameters for $\hat{\gamma}_1$ and $\hat{\gamma}_2$ respectively. For $\hat{\gamma}_2$, we run the same parameters, however we add in a second set of C-S hazard target parameters. The values for both sets of simulations were chosen to mimic realistic scenarios for EHR datasets. We begin by specifying baseline hazards that might be similar to AE risks. For $\hat{\gamma}_1$, we will generate data from two baseline hazards that are selected to represent small and medium baseline risks. Next, we will introduce sets of target regression parameters for each of these baseline hazards. Since we are interested in drug safety, we will consider both the scenario of no drug effect on the log hazard, followed by a slight increase in this relationship. Finally, for both of these scenarios, we can introduce a number of covariate regression parameters to observe these additional effects. Note, that we wish to ensure that adequate amounts of data are generated in each of the $\tilde{\delta}$ groups and as such, we must be careful to select values of ν that are not unduly large, as this may generate a large number of observations that are truly AE occurrences. Hence, we wish to select small values of ν relative to the hazards generated as previously described. We will begin by selecting censorship parameters roughly half of the baseline hazard, and then set them to be equivalent to the baseline hazard. For $\hat{\gamma}_2$, we will retain the same simulation parameters; however we will also include additional parameters for the second C-S hazard. In this case, we will follow similar logic in selection of the hazards for $\hat{\gamma}_1$. However, we will emulate two scenarios. The first will keep the C-S hazards similar in severity, while the second will use target parameters that will create large differences. For the censorship parameter, we will consider the same pattern as before, however we will base this on the all cause baseline hazard.

For the misclassification parameters, we will begin by noting that in the estimation of $\hat{\gamma}_1$, we only have a single parameter to specify, θ. Since, θ represents correct classification, we will select large values and slowly decrease them; specifically $\theta = (0.99, 0.95, 0.9)$. For $\hat{\gamma}_2$, ϕ will take on the same values as θ, but there are two additional misclassification parameters to consider, ξ_1 and ξ_2. As such, we will generate from a subset (for brevity) of all possible combinations of ϕ, ξ_1 and ξ_2, where each can take on any of the three values discussed for $\hat{\gamma}_1$.

Next, we will consider original EHR data sample sizes of 1,000 and 10,000, as EHR databases are usually quite large. For $\hat{\gamma}_1$, we only draw one validation sample and will select 5%, 10% or 20% of the original data sample size. For $\hat{\gamma}_2$ we will consider overall validation sample sizes of 10% and 20% and adjust the values of m_1 and m_0 to observe the effect on estimation. This will be done by starting with $m_1 < m_0$, followed by $m_1 = m_0$ and finally, $m_1 > m_0$. These validation sample sizes will be run for each value of the original sample size, m. To generate the data associated with drug utilization status and the additional binary covariate, z_1, we will draw from a binomial random generator with success probabilities, $P(d)$ and $P(z_1)$. For these values, we will choose $P(d) = 0.4$ and $P(z_1) = 0.6$. Next, for the positive continuous covariate, z_2, we will select a gamma distribution with parameters, $\alpha = 0.5$, $\beta = 1$. This choice reflects the modelling of amounts of a crucial biochemical; perhaps a blood marker.

5.7 Results for Simulation Studies Investigating $L^{(1)}$

First, we will consider the single AE case and run a variety of simulation studies using the target simulation parameters in Table 5.7.1. Parameter set 1 represents a scenario in which there is no additional risk, over the baseline risk of 0.05, associated with drug utilization; however, the log hazard will increase by 0.4 relative to the binary covariate z_1. We have specified a low degree of misclassification by setting $\theta = 0.99$, and demonstrate the effect of slowly decreasing this value in parameter sets 2 and 3. Next, we add in a small drug effect where the log-hazard will increase by 0.1 in the drug utilization group and simultaneously introduce a lessening of risk relative to the continuous covariate, z_2. We will only present the results associated with $\theta = 0.95$ as results for the other cases are similar. So far, we have chosen a value for the censorship parameter that is close to the baseline hazard $\nu = 0.05, 0.14$. Thus, we will introduce a final parameter set where we increase the value of ν past the baseline hazard, which will increase the number of failures or observations with $\delta = 1$. This will be considered in parameter set 5. This case will also have larger covariate effects on the log hazard than the previous sets, as well as a higher baseline hazard.

Table 5.7.1: Target parameter for simulation study investigating univariate model, $L^{(1)}$.

ID	λ_0	β_1	ψ_1	ψ_2	θ	ν
1	0.05	0	0.4	0	0.99	0.05
2	0.05	0	0.4	0	0.95	0.05
3	0.05	0	0.4	0	0.9	0.05
4	0.15	0.1	0.1	-0.25	0.95	0.14
5	0.6	-0.05	0.8	0.5	0.95	1

We display the results for these five parameter sets in the following tables and present the results associated with validation samples of $m = 500, 1000, 2000$ with an original EHR data sample size of $n = 10000$. In Tables 5.7.7 and 5.7.8 we consider parameter set 4 and choose values of $n = 1000, 5000$ to observe the effect of altering the overall EHR data sample size. The validation sample sizes, are fixed at 10%, since this produced good estimates in all other cases. Finally, many other simulations were conducted and produced similar results, enhancing the generalizability of this work. These additional tables are presented in Appendix D.

Table 5.7.2: Results for the simulation study investigating univariate model, $L^{(1)}$, for parameter set 1 with n = 10000.

(a) Summary table for m = 500.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.05	0.04898	0.00154	0.895	0.04995	0.00161	0.0016	0.949
β	0	-9e-05	0.02745	0.937	-0.00017	0.0279	0.02925	0.935
ψ_1	0.4	0.3959	0.02806	0.944	0.40076	0.02848	0.02842	0.948
ψ_2	0	0.00016	0.00951	0.945	0.00021	0.00967	0.00975	0.95
θ	0.99	-	-	-	0.99012	0.00307	0.00306	0.936

(b) Summary table for m =1000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.05	0.04911	0.00154	0.9	0.05009	0.0016	0.00156	0.958
β	0	-3e-04	0.02744	0.95	-0.00036	0.02782	0.02795	0.948
ψ_1	0.4	0.39469	0.02804	0.945	0.39943	0.0284	0.02776	0.953
ψ_2	0	-0.00036	0.00951	0.941	-0.00025	0.00964	0.0097	0.952
θ	0.99	-	-	-	0.9901	0.00245	0.00252	0.931

(c) Summary table for m = 2000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.05	0.04901	0.00154	0.891	0.04999	0.00159	0.00154	0.958
β	0	-0.00104	0.02744	0.934	-0.00071	0.02769	0.02819	0.944
ψ_1	0.4	0.3963	0.02805	0.941	0.40121	0.02827	0.02856	0.949
ψ_2	0	0.00039	0.00951	0.934	0.00038	0.00959	0.01001	0.943
θ	0.99	-	-	-	0.99	0.00188	0.00184	0.942

Table 5.7.3: Results for the simulation study investigating univariate model, $L^{(1)}$, for parameter set 2 with n = 10000.

(a) Summary table for m = 500.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.05	0.04522	0.00145	0.111	0.04999	0.00174	0.00174	0.95
β	0	0.00021	0.02801	0.936	0.00058	0.02962	0.02943	0.952
ψ_1	0.4	0.37719	0.02864	0.846	0.40043	0.03019	0.03082	0.944
ψ_2	0	0.00041	0.00971	0.942	0.00027	0.01027	0.00973	0.961
θ	0.95	-	-	-	0.95001	0.00555	0.00554	0.946

(b) Summary table for m = 1000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.05	0.04526	0.00145	0.12	0.05006	0.00171	0.00168	0.953
β	0	-0.00021	0.02802	0.942	-0.00031	0.02935	0.02898	0.953
ψ_1	0.4	0.37641	0.02863	0.851	0.3997	0.0299	0.03082	0.952
ψ_2	0	4e-05	0.00972	0.927	-1e-05	0.01017	0.01051	0.945
θ	0.95	-	-	-	0.94974	0.00478	0.00478	0.952

(c) Summary table for m = 2000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.05	0.04525	0.00145	0.128	0.05002	0.00166	0.00176	0.937
β	0	-0.00067	0.02801	0.934	-0.00026	0.02879	0.02908	0.948
ψ_1	0.4	0.37756	0.02863	0.868	0.40042	0.02937	0.02939	0.948
ψ_2	0	0.00014	0.00971	0.928	2e-05	0.00998	0.01011	0.951
θ	0.95	-	-	-	0.95011	0.0039	0.00389	0.941

Table 5.7.4: Results for the simulation study investigating univariate model, $L^{(1)}$, for parameter set 3 with n = 10000.

(a) Summary table for m = 500.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.05	0.04093	0.00135	0	0.05005	0.00186	0.00179	0.957
β	0	-0.00092	0.02878	0.914	-0.00063	0.03141	0.03171	0.941
ψ_1	0.4	0.35706	0.02942	0.665	0.40022	0.032	0.03237	0.946
ψ_2	0	0.00016	0.00997	0.928	0.00014	0.01088	0.01052	0.96
θ	0.9	-	-	-	0.90001	0.00693	0.007	0.949

(b) Summary table for m = 1000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.05	0.04098	0.00135	0	0.05003	0.00181	0.00176	0.956
β	0	4e-05	0.02879	0.928	0.00043	0.03092	0.0311	0.951
ψ_1	0.4	0.35546	0.02942	0.651	0.39983	0.0315	0.03158	0.951
ψ_2	0	-0.00043	0.00998	0.918	-0.00034	0.01072	0.01064	0.95
θ	0.9	-	-	-	0.90006	0.00615	0.00613	0.953

(c) Summary table for m = 2000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.05	0.04097	0.00135	0	0.05007	0.00173	0.00177	0.947
β	0	-1e-04	0.02879	0.937	-0.00067	0.02999	0.02904	0.956
ψ_1	0.4	0.35639	0.02942	0.68	0.40075	0.03057	0.0294	0.964
ψ_2	0	-0.00018	0.00997	0.911	-3e-04	0.01039	0.01044	0.964
θ	0.9	-	-	-	0.9001	0.0052	0.00516	0.95

Table 5.7.5: Results for the simulation study investigating univariate model, $L^{(1)}$, for parameter set 4 with n = 10000.

(a) Summary table for m = 500.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.15	0.13587	0.00496	0.212	0.15032	0.00605	0.00622	0.952
β	0.1	0.0962	0.03211	0.943	0.10013	0.03349	0.03271	0.946
ψ_1	0.1	0.09481	0.03246	0.937	0.09837	0.03383	0.03409	0.945
ψ_2	-0.25	-0.24216	0.01306	0.895	-0.25036	0.0135	0.01358	0.943
θ	0.95	-	-	-	0.95013	0.00834	0.00836	0.942

(b) Summary table for m = 1000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.15	0.13556	0.00495	0.203	0.14978	0.00584	0.00626	0.938
β	0.1	0.09604	0.03212	0.95	0.0998	0.03329	0.03287	0.954
ψ_1	0.1	0.09746	0.03248	0.94	0.10138	0.03363	0.03491	0.948
ψ_2	-0.25	-0.24176	0.01306	0.883	-0.24963	0.01342	0.01377	0.94
θ	0.95	-	-	-	0.95035	0.00676	0.00651	0.955

(c) Summary table for m = 2000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.15	0.13604	0.00497	0.222	0.15043	0.00569	0.00557	0.953
β	0.1	0.09633	0.03211	0.937	0.10012	0.03289	0.03372	0.947
ψ_1	0.1	0.09628	0.03247	0.95	0.09994	0.03323	0.03305	0.954
ψ_2	-0.25	-0.24326	0.01307	0.901	-0.25125	0.0133	0.01332	0.952
θ	0.95	-	-	-	0.95007	0.00524	0.0053	0.948

Table 5.7.6: Results for the simulation study investigating univariate model, $L^{(1)}$, for parameter set 5 with n = 10000.

(a) Summary table for m = 500.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.6	0.65808	0.01902	0.337	0.60058	0.01917	0.01889	0.958
β	-0.05	-0.04285	0.02518	0.764	-0.05293	0.02645	0.02676	0.949
ψ_1	0.8	0.68362	0.02592	0.058	0.80033	0.02725	0.02779	0.943
ψ_2	0.5	0.38435	0.00773	0.001	0.50019	0.00851	0.00822	0.958
θ	0.95	-	-	-	0.95002	0.00357	0.00368	0.936

(b) Summary table for m = 1000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.6	0.65772	0.01904	0.341	0.60016	0.01881	0.0189	0.956
β	-0.05	-0.0401	0.02518	0.747	-0.04953	0.02614	0.02537	0.952
ψ_1	0.8	0.68461	0.02591	0.055	0.80093	0.02692	0.0281	0.933
ψ_2	0.5	0.38368	0.00774	0	0.49976	0.00843	0.00823	0.962
θ	0.95	-	-	-	0.94999	0.00331	0.00339	0.943

(c) Summary table for m = 2000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.6	0.65848	0.01904	0.342	0.60074	0.01828	0.01767	0.958
β	-0.05	-0.04094	0.02517	0.755	-0.04914	0.02556	0.02407	0.967
ψ_1	0.8	0.68355	0.0259	0.051	0.80021	0.02631	0.02721	0.945
ψ_2	0.5	0.38421	0.00772	0	0.49996	0.0083	0.00858	0.937
θ	0.95	-	-	-	0.94989	0.00294	0.00297	0.953

Table 5.7.7: Results for the simulation study investigating univariate model, $L^{(1)}$, for parameter set 4 with n = 1000, m = 100.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.15	0.1375	0.01601	0.816	0.15144	0.01889	0.01915	0.942
β	0.1	0.09894	0.10186	0.952	0.10186	0.10553	0.10316	0.964
ψ_1	0.1	0.09439	0.10296	0.946	0.09929	0.10653	0.10354	0.958
ψ_2	-0.25	-0.24472	0.04154	0.935	-0.25227	0.04265	0.04318	0.946
θ	0.95	-	-	-	0.95215	0.02174	0.02238	0.917

Table 5.7.8: Results for the simulation study investigating univariate model, $L^{(1)}$, for parameter set 4 with n = 5000, m = 500.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}_1$	$\hat{\sigma}(\hat{\gamma}_1)$	s	$\widehat{CP}$
λ	0.15	0.1357	0.00702	0.474	0.14998	0.00829	0.00814	0.958
β	0.1	0.09832	0.04546	0.945	0.10232	0.04711	0.0458	0.948
ψ_1	0.1	0.09565	0.04598	0.951	0.09937	0.04759	0.04738	0.956
ψ_2	-0.25	-0.2428	0.0185	0.927	-0.25081	0.01901	0.01888	0.948
θ	0.95	-	-	-	0.95036	0.00953	0.00987	0.931

Each of the tables of results first presents the target parameter values as described in Table 5.7.1. The columns marked $\tilde{\gamma}$, $\hat{\sigma}(\tilde{\gamma})$ and $\widetilde{CP}$ denote the results ignoring the presence of misclassification and are produced by the *survreg*() function in R. The columns marked $\hat{\gamma}_1$, $\hat{\sigma}(\hat{\gamma}_1)$, s and $\widehat{CP}$ are the results associated with the likelihood $L^{(1)}$. The column marked 's' denotes the sample standard deviation of the estimates of γ the CP columns denote the proportion that the target parameter was housed in a 95% confidence interval. Note, that these columns are the same for the following section describing the competing risks results.

The results presented demonstrate that the MCB-adjusted MLE $\hat{\gamma}_1$ performs well in all cases presented here. This is also true of the additional simulations included in Appendix D. The results for parameter sets 1-3 are all more or less the same for $\tilde{\gamma}_1$; however we note slightly low coverage for the baseline hazard. As the misclassification rate increases, these estimates worsen with respect to bias and coverage as is noted in Tables 5.7.3 and 5.7.4. However, the MCB-adjusted estimator continues to perform well even at the lowest validation sample size of $m = 500$. For parameter sets 4 and 5 we continue to see poor coverage estimates for $\tilde{\gamma}$, whereas our estimator continues to perform well. Finally, in Table 5.7.7 there is a slight drop off in coverage associated with the θ-parameter. This demonstrates the problematic effect of smaller sample sizes. However, the loss in coverage is not large and in Table 5.7.8, with $n = 5000$, $m = 500$, the coverage estimates improve.

Thus, it appears that $\hat{\gamma}_1$ does consistently well at adjusting for misclassification of the outcome in a single AE scenario. Next, we will conduct a similar set of simulation studies to observe the behaviour of $\hat{\gamma}_2$.

5.8 Results for Simulation Studies Investigating $L^{(2)}$

The selection of the target simulation parameters follows from the parameter sets in the previous section. However, we will select additional parameters for the competing C-S hazard. For the first two parameters sets, we will use parameter sets 1 and 3 from the previous section (Table 5.7.1). This will allow us to observe the increase in misclassification from 1% to 10% with respect to ϕ. In addition, the alternate AE will have a slightly higher baseline hazard, λ_{02}, and will additionally have an increase in the log-hazard associated with z_2, as can be seen by the η_2 column. We introduce additional misclassification rates associated with the competing event type of 90% for correctly classifying an AE of type 1, ξ_1, and 95% for type 2, ξ_2. Next, we will consider the target parameters in parameter set 4 from Table 5.7.1. For the second C-S hazard we will consider a case in which the baseline hazard is much higher and we have a larger increase on the log-hazard associated with drug utilization. We also consider a decreasing log-hazard based on the additional covariates. Next, we will switch the target values for the ξ-parameters and consider both $\phi = 0.9, 0.99$. To investigate the increase of the censorship parameter, we will reconsider parameter set 2 of Table 5.7.1; however, we will increase the value of ν.

Table 5.8.1: Target parameters for simulation study investigating competing risks model, $L^{(2)}$.

ID	λ_{01}	β_1	ψ_1	ψ_2	λ_{02}	β_1	η_1	η_2	ϕ	ξ_1	ξ_2	ν
1	0.05	0	0.4	0	0.06	0	0.2	0.5	0.99	0.9	0.95	0.055
2	0.05	0	0.4	0	0.06	0	0.2	0.5	0.9	0.9	0.95	0.055
3	0.15	0.1	0.1	-0.25	0.4	0.5	-0.15	-0.5	0.99	0.95	0.9	0.25
4	0.15	0.1	0.1	-0.25	0.4	0.5	-0.15	-0.5	0.9	0.95	0.9	0.25
5	0.05	0	0.4	0	0.06	0	0.2	0.5	0.9	0.9	0.95	0.11

As in the previous section, we will consider a sample size of $n = 10000$ and three validation sample sizes, $m = 500, 1000, 2000$. We will also consider lowering n to observe the effect in a similar manner as in the previous section by considering $n = 1000$ and 5000; however, we will use a larger validation sample size at 20%. Finally, we are drawing two validation samples, one from each $V_{\tilde{\delta}}$, $\tilde{\delta} = 0, 1$. To begin we will consider drawing 50% from each validation subgroup, followed by altering the proportion of observations in the validation sample drawn from V_1 to observe the effect. First, we will select $m_1 = 0.3 \times m$ followed by $m_1 = 0.7 \times m$ where $m_0 = m - m_1$. Since the results are similar we present only those associated with simulation parameter set 4 with sample sizes $n = 10000$ and $m = 1000$. Again, many simulation studies were conducted and the results were omitted from this discussion as they are all similar in nature; these additional results are displayed in Appendix D.

Table 5.8.2: Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 1 with n = 10000.

(a) Summary Tables: m = 500.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.05	0.07095	0.0022	0	0.05004	0.00263	0.00262	0.959
β_1	0	0.00138	0.0316	0.935	0.00126	0.04081	0.04073	0.954
ψ_1	0.4	0.29865	0.032	0.12	0.39919	0.04219	0.04058	0.964
ψ_2	0	-0.04481	0.0287	0.662	-0.00195	0.04253	0.04433	0.947
λ_2	0.06	0.08938	0.0022	0	0.05992	0.00257	0.00258	0.951
β_2	0	-0.00046	0.0268	0.923	-0.00089	0.03434	0.03593	0.94
η_1	0.2	0.12698	0.0266	0.227	0.20008	0.03453	0.03391	0.953
η_2	0.5	0.39812	0.0167	0	0.50052	0.02043	0.02061	0.949
ϕ	0.99	-	-	-	0.99002	0.00197	0.00196	0.949
ξ_1	0.9	-	-	-	0.9012	0.02605	0.02701	0.924
ξ_2	0.95	-	-	-	0.95056	0.01641	0.01688	0.92

(b) Summary Tables: m = 1000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.05	0.07089	0.0022	0	0.04995	0.0023	0.0023	0.942
β_1	0	-1e-05	0.0316	0.938	-7e-05	0.04055	0.0402	0.956
ψ_1	0.4	0.30059	0.032	0.138	0.40084	0.04175	0.0414	0.956
ψ_2	0	-0.0443	0.0287	0.658	-8e-05	0.03957	0.04025	0.941
λ_2	0.06	0.0895	0.0022	0	0.06005	0.00224	0.00215	0.953
β_2	0	0.00054	0.0267	0.931	0.00011	0.03409	0.03452	0.948
η_1	0.2	0.12728	0.0266	0.228	0.2003	0.03415	0.03374	0.951
η_2	0.5	0.39774	0.0167	0	0.49985	0.01965	0.01965	0.955
ϕ	0.99	-	-	-	0.98992	0.00181	0.00183	0.949
ξ_1	0.9	-	-	-	0.90164	0.01869	0.01813	0.943
ξ_2	0.95	-	-	-	0.95101	0.01173	0.01215	0.918

(c) Summary Tables: m =2000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.05	0.07083	0.0022	0	0.04994	0.00211	0.0021	0.953
β_1	0	-0.00088	0.0316	0.943	-0.00118	0.04016	0.04058	0.954
ψ_1	0.4	0.30116	0.032	0.136	0.40198	0.04125	0.04077	0.957
ψ_2	0	-0.04219	0.0286	0.672	0.00164	0.03781	0.03881	0.949
λ_2	0.06	0.08956	0.0022	0	0.06	0.00205	0.00202	0.949
β_2	0	8e-05	0.0267	0.929	0.00052	0.03379	0.03537	0.938
η_1	0.2	0.12611	0.0266	0.211	0.20031	0.03379	0.03246	0.962
η_2	0.5	0.39782	0.0167	0	0.50104	0.01919	0.01942	0.949
ϕ	0.99	-	-	-	0.98999	0.00157	0.00157	0.947
ξ_1	0.9	-	-	-	0.90003	0.01344	0.01317	0.953
ξ_2	0.95	-	-	-	0.95019	0.00848	0.00855	0.944

Table 5.8.3: Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 2 with n = 10000.

(a) Summary Tables: m = 500.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.05	0.06103	0.0018	0	0.05007	0.00272	0.00273	0.945
β_1	0	-0.00212	0.0316	0.904	0.00109	0.04387	0.04594	0.94
ψ_1	0.4	0.24838	0.032	0.009	0.39858	0.04527	0.04405	0.966
ψ_2	0	-0.1209	0.0284	0.019	4e-04	0.04415	0.04546	0.946
λ_2	0.06	0.07723	0.0019	0	0.05991	0.00268	0.00279	0.94
β_2	0	0	0.0268	0.888	0.00262	0.03723	0.03491	0.965
η_1	0.2	0.07142	0.0266	0.017	0.19982	0.03745	0.03787	0.952
η_2	0.5	0.31063	0.0163	0	0.50017	0.0218	0.02214	0.941
ϕ	0.9	-	-	-	0.90003	0.00469	0.00463	0.953
ξ_1	0.9	-	-	-	0.90044	0.02592	0.02647	0.929
ξ_2	0.95	-	-	-	0.9509	0.01616	0.01574	0.922

(b) Summary Tables: m = 1000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.05	0.06095	0.0018	0	0.05002	0.00242	0.00228	0.971
β_1	0	-0.00088	0.0316	0.91	-0.00018	0.0434	0.0435	0.946
ψ_1	0.4	0.2508	0.032	0.005	0.40022	0.04465	0.04454	0.951
ψ_2	0	-0.12201	0.0284	0.021	-0.00159	0.04169	0.04146	0.953
λ_2	0.06	0.07725	0.0019	0	0.06007	0.00237	0.00241	0.944
β_2	0	-0.00091	0.0268	0.87	-0.00113	0.0368	0.03738	0.95
η_1	0.2	0.07301	0.0266	0.013	0.19883	0.03689	0.03774	0.951
η_2	0.5	0.30951	0.0163	0	0.50066	0.02104	0.02162	0.938
ϕ	0.9	-	-	-	0.90001	0.00452	0.00434	0.956
ξ_1	0.9	-	-	-	0.89991	0.01863	0.01836	0.946
ξ_2	0.95	-	-	-	0.94968	0.01182	0.0121	0.936

(c) Summary Tables: m = 2000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.05	0.06095	0.0018	0	0.04999	0.0022	0.00215	0.955
β_1	0	0.00015	0.0316	0.916	0.00035	0.04237	0.04246	0.948
ψ_1	0.4	0.24993	0.032	0.01	0.40042	0.04349	0.04224	0.953
ψ_2	0	-0.12134	0.0284	0.03	-6e-04	0.03936	0.03958	0.951
λ_2	0.06	0.07721	0.0019	0	0.06006	0.00216	0.00216	0.945
β_2	0	-0.00038	0.0268	0.875	0.00061	0.03598	0.03495	0.959
η_1	0.2	0.07264	0.0266	0.01	0.19877	0.03599	0.03529	0.951
η_2	0.5	0.30916	0.0162	0	0.49996	0.02042	0.02031	0.947
ϕ	0.9	-	-	-	0.89995	0.00425	0.00448	0.939
ξ_1	0.9	-	-	-	0.90062	0.01332	0.01331	0.945
ξ_2	0.95	-	-	-	0.9499	0.00847	0.0083	0.948

Table 5.8.4: Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 3 with n = 10000.

(a) Summary Tables: m = 500.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.15	0.24279	0.0081	0	0.15077	0.01191	0.01208	0.947
β_1	0.1	0.03654	0.0362	0.574	0.09878	0.06025	0.05904	0.954
ψ_1	0.1	0.08793	0.0358	0.934	0.09966	0.05741	0.05772	0.955
ψ_2	-0.25	-0.10144	0.0222	0	-0.25142	0.03932	0.03981	0.95
λ_2	0.4	0.53514	0.0126	0	0.39997	0.01524	0.01499	0.95
β_2	0.5	0.35503	0.0248	0	0.50035	0.03235	0.03212	0.948
η_1	-0.15	-0.11803	0.0249	0.743	-0.14961	0.03238	0.03142	0.952
η_2	-0.5	-0.31214	0.0191	0	-0.50088	0.02896	0.02771	0.958
ϕ	0.99	-	-	-	0.98993	0.00243	0.00245	0.945
ξ_1	0.95	-	-	-	0.95033	0.0272	0.02387	0.911
ξ_2	0.9	-	-	-	0.90117	0.01892	0.01934	0.926

(b) Summary Tables: m = 1000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.15	0.24295	0.0081	0	0.15005	0.00981	0.00961	0.957
β_1	0.1	0.03488	0.0362	0.573	0.0951	0.05763	0.05783	0.949
ψ_1	0.1	0.08757	0.0358	0.947	0.09953	0.05609	0.05418	0.958
ψ_2	-0.25	-0.10082	0.0222	0	-0.25017	0.03839	0.03721	0.952
λ_2	0.4	0.53544	0.0126	0	0.40054	0.01365	0.01384	0.949
β_2	0.5	0.35495	0.0248	0	0.50018	0.03193	0.03212	0.952
η_1	-0.15	-0.11791	0.0249	0.743	-0.14881	0.0321	0.03168	0.954
η_2	-0.5	-0.31333	0.0192	0	-0.50093	0.02862	0.02847	0.946
ϕ	0.99	-	-	-	0.99	0.00218	0.00221	0.936
ξ_1	0.95	-	-	-	0.95074	0.01726	0.01718	0.917
ξ_2	0.9	-	-	-	0.89947	0.01355	0.01375	0.94

(c) Summary Tables: m = 2000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.15	0.24287	0.0081	0	0.15024	0.00855	0.008	0.959
β_1	0.1	0.03571	0.0362	0.586	0.09643	0.0553	0.05639	0.948
ψ_1	0.1	0.08861	0.0358	0.94	0.10267	0.05456	0.05267	0.955
ψ_2	-0.25	-0.10085	0.0223	0	-0.25022	0.03753	0.03644	0.957
λ_2	0.4	0.53572	0.0126	0	0.39984	0.01269	0.01234	0.96
β_2	0.5	0.35498	0.0248	0	0.50093	0.03155	0.03085	0.958
η_1	-0.15	-0.11867	0.0249	0.759	-0.14879	0.03177	0.03293	0.95
η_2	-0.5	-0.3116	0.0192	0	-0.49968	0.02834	0.02874	0.944
ϕ	0.99	-	-	-	0.98996	0.00186	0.00196	0.93
ξ_1	0.95	-	-	-	0.94932	0.01241	0.01259	0.932
ξ_2	0.9	-	-	-	0.90022	0.00975	0.00974	0.945

Table 5.8.5: Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 4 with n = 10000.

(a) Summary Tables: m = 500.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.15	0.20162	0.0068	0	0.1499	0.01204	0.01211	0.948
β_1	0.1	-0.02476	0.0364	0.079	0.09924	0.0634	0.06211	0.962
ψ_1	0.1	0.10158	0.036	0.913	0.10232	0.06098	0.06318	0.936
ψ_2	-0.25	-0.05254	0.0221	0	-0.24918	0.04179	0.04161	0.952
λ_2	0.4	0.45123	0.0106	0.004	0.39999	0.01623	0.01664	0.945
β_2	0.5	0.29492	0.0248	0	0.49906	0.03549	0.03717	0.942
η_1	-0.15	-0.10691	0.0248	0.57	-0.14904	0.03556	0.03489	0.958
η_2	-0.5	-0.26429	0.019	0	-0.49991	0.0315	0.03151	0.955
ϕ	0.9	-	-	-	0.90005	0.00547	0.00558	0.949
ξ_1	0.95	-	-	-	0.95115	0.027	0.02352	0.914
ξ_2	0.9	-	-	-	0.90008	0.01872	0.01849	0.944

(b) Summary Tables: m = 1000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.15	0.20154	0.0068	0	0.14959	0.01003	0.01033	0.943
β_1	0.1	-0.02512	0.0364	0.102	0.10253	0.06032	0.06164	0.945
ψ_1	0.1	0.10057	0.036	0.929	0.10135	0.05915	0.06053	0.949
ψ_2	-0.25	-0.0517	0.0221	0	-0.24881	0.04059	0.04105	0.957
λ_2	0.4	0.45125	0.0106	0.008	0.40043	0.0147	0.01599	0.928
β_2	0.5	0.29701	0.0248	0	0.50063	0.03485	0.03522	0.944
η_1	-0.15	-0.10563	0.0248	0.565	-0.14955	0.03505	0.03615	0.941
η_2	-0.5	-0.26445	0.019	0	-0.5015	0.03106	0.03012	0.96
ϕ	0.9	-	-	-	0.90022	0.00522	0.00542	0.944
ξ_1	0.95	-	-	-	0.95036	0.01733	0.01741	0.922
ξ_2	0.9	-	-	-	0.9	0.0134	0.01355	0.94

(c) Summary Tables: m =2000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.15	0.20155	0.0068	0	0.1499	0.00877	0.00862	0.95
β_1	0.1	-0.02639	0.0364	0.078	0.10005	0.05767	0.05584	0.962
ψ_1	0.1	0.10291	0.036	0.935	0.09958	0.05705	0.05558	0.951
ψ_2	-0.25	-0.05397	0.0221	0	-0.25214	0.03935	0.04018	0.943
λ_2	0.4	0.45114	0.0106	0	0.40011	0.01363	0.01358	0.948
β_2	0.5	0.29547	0.0248	0	0.50019	0.03401	0.03385	0.952
η_1	-0.15	-0.10527	0.0248	0.551	-0.14971	0.03425	0.03326	0.957
η_2	-0.5	-0.26422	0.019	0	-0.5012	0.03026	0.03	0.952
ϕ	0.9	-	-	-	0.90027	0.00482	0.00476	0.953
ξ_1	0.95	-	-	-	0.95003	0.01233	0.0122	0.938
ξ_2	0.9	-	-	-	0.90001	0.00968	0.00976	0.951

Table 5.8.6: Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 5 with n = 10000.

(a) Summary Tables: m = 500.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.05	0.08884	0.0027	0	0.05001	0.003	0.00288	0.956
β_1	0	-0.00194	0.0315	0.928	-0.00143	0.04951	0.05035	0.948
ψ_1	0.4	0.23287	0.0319	0.002	0.40179	0.05131	0.05037	0.956
ψ_2	0	-0.13816	0.0286	0.005	-0.00209	0.04787	0.04705	0.954
λ_2	0.06	0.11088	0.0028	0	0.05998	0.00296	0.003	0.953
β_2	0	-0.00095	0.0268	0.915	-0.00109	0.04191	0.04074	0.953
η_1	0.2	0.0551	0.0266	0	0.20049	0.04223	0.04242	0.952
η_2	0.5	0.30791	0.0164	0	0.50033	0.02356	0.02411	0.943
ϕ	0.9	-	-	-	0.90001	0.00698	0.00683	0.953
ξ_1	0.9	-	-	-	0.90091	0.02606	0.0273	0.924
ξ_2	0.95	-	-	-	0.95087	0.01627	0.01614	0.925

(b) Summary Tables: m = 1000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.05	0.08874	0.0027	0	0.04993	0.0027	0.00261	0.951
β_1	0	0.00118	0.0315	0.941	0.00323	0.04903	0.0503	0.939
ψ_1	0.4	0.23173	0.0319	0.001	0.39726	0.05062	0.04875	0.964
ψ_2	0	-0.13502	0.0286	0.006	-0.00184	0.04543	0.04485	0.949
λ_2	0.06	0.11054	0.0028	0	0.05996	0.00266	0.00268	0.945
β_2	0	0.0026	0.0268	0.918	0.00166	0.04142	0.04116	0.944
η_1	0.2	0.0556	0.0266	0.001	0.20266	0.04162	0.0424	0.945
η_2	0.5	0.31016	0.0164	0	0.50059	0.02276	0.02305	0.949
ϕ	0.9	-	-	-	0.90017	0.0065	0.00653	0.955
ξ_1	0.9	-	-	-	0.90103	0.01875	0.01882	0.943
ξ_2	0.95	-	-	-	0.94928	0.0119	0.01207	0.938

(c) Summary Tables: m = 2000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.05	0.08882	0.0027	0	0.04995	0.00249	0.00243	0.969
β_1	0	0.00193	0.0315	0.93	0.00187	0.04786	0.05004	0.935
ψ_1	0.4	0.23204	0.0319	0	0.40269	0.04939	0.04778	0.956
ψ_2	0	-0.13731	0.0286	0	0.00147	0.04284	0.04343	0.945
λ_2	0.06	0.11097	0.0028	0	0.05997	0.00245	0.00243	0.944
β_2	0	0.00069	0.0268	0.933	0.00278	0.04062	0.04041	0.948
η_1	0.2	0.0526	0.0266	0	0.19985	0.04077	0.0406	0.947
η_2	0.5	0.30807	0.0164	0	0.50035	0.02217	0.02167	0.95
ϕ	0.9	-	-	-	0.9	0.00583	0.00595	0.94
ξ_1	0.9	-	-	-	0.9003	0.01348	0.0135	0.945
ξ_2	0.95	-	-	-	0.95064	0.00846	0.00813	0.949

Table 5.8.7: Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 5 with n = 1000, m = 200.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.05	0.0713	0.0069	0	0.05031	0.00682	0.00664	0.957
β_1	0	-0.00184	0.1003	0.013	-0.00036	0.12782	0.12879	0.952
ψ_1	0.4	0.30381	0.10154	0.157	0.40635	0.13174	0.12939	0.952
ψ_2	0	-0.0455	0.09164	0.809	-0.01026	0.12481	0.1259	0.954
λ_2	0.06	0.08974	0.0071	0	0.06024	0.00661	0.00676	0.947
β_2	0	-0.00024	0.08487	0.344	0.00328	0.10774	0.1096	0.947
η_1	0.2	0.12263	0.08432	0.013	0.19455	0.10782	0.114	0.924
η_2	0.5	0.40485	0.05361	0	0.50311	0.06261	0.06237	0.945
ϕ	0.99	-	-	-	0.99	0.00531	0.00495	0.921
ξ_1	0.9	-	-	-	0.90296	0.04222	0.04155	0.919
ξ_2	0.95	-	-	-	0.95164	0.03191	0.02647	0.938

Table 5.8.8: Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 5 with n = 5000, m = 1000.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.05	0.0711	0.00306	0	0.05007	0.00299	0.003	0.957
β_1	0	0.00052	0.04476	0	0.00044	0.0568	0.05566	0.97
ψ_1	0.4	0.29742	0.04528	0	0.39743	0.05836	0.05948	0.944
ψ_2	0	-0.04493	0.04061	0.28	0.00032	0.05353	0.05219	0.95
λ_2	0.06	0.08962	0.00315	0	0.06004	0.0029	0.00296	0.947
β_2	0	-0.00115	0.03782	0.002	-0.00087	0.04781	0.04822	0.951
η_1	0.2	0.12719	0.03758	0	0.2012	0.04781	0.04603	0.964
η_2	0.5	0.39786	0.02362	0	0.50044	0.02723	0.02759	0.958
ϕ	0.99	-	-	-	0.99002	0.00222	0.00228	0.946
ξ_1	0.9	-	-	-	0.90122	0.01889	0.01898	0.936
ξ_2	0.95	-	-	-	0.95059	0.01191	0.0123	0.93

Table 5.8.9: Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 4 with n = 10000, m1 = 300, m0 = 700.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.15	0.20181	0.0068	0	0.14988	0.01111	0.01095	0.954
β_1	0.1	-0.02693	0.03637	0.067	0.09707	0.06191	0.0592	0.96
ψ_1	0.1	0.10186	0.03598	0	0.10015	0.05988	0.06027	0.953
ψ_2	-0.25	-0.05389	0.02206	0	-0.25155	0.0411	0.04117	0.951
λ_2	0.4	0.45208	0.01058	0.025	0.40106	0.01537	0.01531	0.951
β_2	0.5	0.29608	0.02478	0	0.50047	0.03484	0.03512	0.944
η_1	-0.15	-0.10778	0.02485	0	-0.15067	0.03497	0.03574	0.946
η_2	-0.5	-0.26541	0.01898	0	-0.50153	0.03099	0.03097	0.956
ϕ	0.9	-	-	-	0.90023	0.00504	0.00487	0.952
ξ_1	0.95	-	-	-	0.95102	0.02337	0.02222	0.918
ξ_2	0.9	-	-	-	0.89974	0.01676	0.01756	0.935

Table 5.8.10: Results for the simulation study investigating competing risks model, $L^{(2)}$, for parameter set 4 with n = 10000, m1 = 700, m0 = 300.

	Target	$\tilde{\gamma}$	$\hat{\sigma}(\tilde{\gamma})$	$\widetilde{CP}$	$\hat{\gamma}$	$\hat{\sigma}(\hat{\gamma})$	s	$\widehat{CP}$
λ_1	0.15	0.20174	0.0068	0	0.15079	0.0111	0.01146	0.94
β_1	0.1	-0.02623	0.03638	0.075	0.09864	0.06169	0.06126	0.948
ψ_1	0.1	0.1015	0.03597	0	0.09882	0.05966	0.05884	0.954
ψ_2	-0.25	-0.05292	0.02207	0	-0.25249	0.04104	0.04195	0.944
λ_2	0.4	0.45114	0.01055	0.02	0.39971	0.01532	0.01526	0.956
β_2	0.5	0.29725	0.02477	0	0.50165	0.03486	0.03409	0.961
η_1	-0.15	-0.10594	0.02484	0	-0.14924	0.03499	0.03644	0.934
η_2	-0.5	-0.26407	0.01898	0	-0.49854	0.03097	0.03072	0.948
ϕ	0.9	-	-	-	0.8999	0.00505	0.00502	0.946
ξ_1	0.95	-	-	-	0.94968	0.0231	0.02241	0.927
ξ_2	0.9	-	-	-	0.90046	0.0167	0.01727	0.946

The results demonstrate that the MCB-adjusted estimator, $\hat{\gamma}_2$ performs quite well under most of the simulation scenarios considered. With $m = 500$ observations, we see a somewhat low (≈ 0.91) estimate of coverage probability; however, increasing to $m = 1000$ appears to resolve the problem. Sample size issues are further noted in Table 5.8.7 where we see low estimates of coverage probability when the misclassification rates that are low; however, results improve when the sample size is increased to $n = 5000$ (Table 5.8.8) demonstrating the effect of lower original sample sizes. Recall, that EHR databases are generally quite large, so that this is not as much of a concern for the motivating example of applications to pharmacovigilance. Further, as the sample sizes increase, the estimates of standard deviation based on the inverse of the observed information matrix approximate the sample standard deviation based on the MLEs. Considering the 'tilde' result columns, we can see that the usual approach to estimation, assuming perfect classification, will clearly produce problematic risk estimates. Finally, when altering the proportion of observations in the validation sample drawn from V_1, we observe little difference in the coverage estimates. This seems reasonable since the contribution to the likelihood from V_1 or V_0 is representing a different form of misclassification and they are mutually exclusive. An observed AE occurrence is assumed to occur without error and is required for a misclassification of the C-S AE type. Alternatively, an observation in which the time-to-event is observed with error is recorded as a LTF event, which implies that the C-S AE type is irrelevant. Thus, alteration of the sample size in either V_1 or V_0 will likely only have an effect on estimation as a result of small sample size.

5.9 Conclusions and Discussion

We have demonstrated the ability for MCB-adjustment in a time-to-event setting under assumptions motivated by EHR data. First, we utilized a parametric model for right censored survival data, assuming a constant baseline hazard and time independent covariates and misclassification rates. The example employed throughout this discussion is based on an exponential time to AE distribution. Next, we assumed the presence of competing AEs under similar assumptions, and provided MCB-adjusted estimation using a parametric C-S hazards formulation of competing risks (Kalbfleisch and Prentice (2002) [19]). Finally, we assumed misclassification is only possible given the true event is an AE occurrence, which is a realistic assumption in the case of EHR data. However, relaxing these assumptions can provide a more general framework for application of the methods proposed here.

First, the use of a partial likelihood approach for either model would require some thought as the baseline hazard does not readily factor out of $P(\tilde{T} \in I_{dt})$. However, the original EHR data portion of $L^{(2b)}$, equation (5.11), is the parametric form of a likelihood originally presented as a semi-parametric model with known misclassification rates by Rompaye et al. (2010) [36]. We can consider the formulation of a validation sample portion to the partial likelihood to estimate the unknown misclassification rates if we assume that the time-to-event is observed without error, and misclassification can only occur based on misspecification of the C-S AE occurrence type. Rompaye et al. (2010) [36] invoke the additional assumption of proportional baseline hazards, which allowed them to factor the partial likelihood.

This assumption introduces another parameter which, if unknown, may introduce unidentifiability in the model, even with the presence of validation data. Both ξ-parameters, the misclassification rates associated with C-S hazards, will depend on the value of this parameter and vice versa. Further, for our purposes, prior knowledge of this quantity is not likely to be available for EHR data. Hence, we did not include this in the main discussion. Future research may be focused on estimating this value in order to derive a partial likelihood for MCB-adjustment within a semi-parametric framework.

Next, we can consider right censorship as a assumption that does not account for all censorship type occurrences in EHR data. Outpatient care is omitted from certain EHR datasets, which can be interpreted as interval censorship. This type of censorship is problematic in scenarios for which the vast majority of treatments are applied out of the hospital setting and we are unable to observe AE occurrence rates. Thus, adjusting these likelihoods to account for this possibility could be a useful addition. Specifically, if we were able to attain validation information for a subset of those provided with outpatient care, we may be able to extend this technique to address this problem.

Relaxing the assumption that misclassification can only occur given a true AE occurrence is unrealistic with EHR data; however, can be considered in other contexts. In order to fail to observe an end point as a censorship, we must have incorrectly observed an event of interest prior to this time point. Continuing to collect data after this erroneously-recorded event could allow for the use of validation data to adjust the likelihood for these errors. In other words, modelling the possibility of recurrent events may facilitate this extension, as we would be interested in observing individuals throughout their entire time in the observational interval, not just the time until first event occurrence.

Finally, modelling the hazard as a function of time dependent covariates can be introduced with the presence of outcome misclassification. Each individual scenario must be characterized carefully since the hazard will depend on both the misclassification rate as well as these covariates. For example, consider dividing the observational interval into two sections, and define a time dependent covariate such that it is fixed within these intervals but varies between these intervals. We can represent this in the likelihood using an indicator variable to designate the membership to that time interval, and the hazard will be adjusted accordingly. However, even with fixed misclassification rates over time, we must account for additional hypothetical situations that can arise under this example in the likelihood. For instance, we observe a misclassified LTF-type event that occurs in time interval two. There are now two possible situations that must appear in the validation sample portion of the likelihood; namely, the true event occurred in either time interval given that we observed an LTF event in the second interval. Thus additional information will be necessary to distinguish between these possibilities. Incorporation of a misclassification rate that depends on time will further complicate estimation. Future work is needed to address these complexities in turn.

Appendix A

Mathematical Results

This discussion focuses on maximum likelihood inference for the additional information housed in a secondary or tertiary internal validation sample. Thus, in this preliminary section, we will review and investigate the asymptotic results associated with the analysis of EHR data under a multi-sample framework. Section A.1 will introduce, without proof but referenced, some standard results needed to develop the relevant asymptotic properties. Section A.2 will expand these to the multi-sample setting. Proofs will be provided or referenced as needed.

A.1 Preliminary Results

For this section, let $\{X_n\}$ and $\{Y_n\}$ denote sequences of random variables. Further we will use $\gamma = (\gamma_1, ..., \gamma_p)^T$ to denote the p-dimensional parameter vector of interest and $\hat{\gamma} = (\hat{\gamma}_1, ..., \hat{\gamma}_p)^T$ to denote it's estimate.

Theorem A.1.1 - Slutsky's Theorem If $X_n \overset{D}{\to} X$ and $Y_n \overset{p}{\to} a$, where a is a constant, then,

i) $Y_n X_n \overset{D}{\to} aX$.

ii) $X_n + Y_n \overset{D}{\to} X + a$.

(*Taken from Casella and Berger Theorem 5.5.17*)

Theorem A.1.2 Suppose $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$. Then, $X_nY_n \xrightarrow{p} XY$.

(*Taken from Hogg McKean and Craig 6th Edition, Theorem 4.2.5*)

Theorem A.1.3 Weak Law of Large Numbers Let $X_1, ..., X_n$ be *iid* random variables with $E(X_i) = \mu$ and $Var(X_i) < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1. \tag{A.1}$$

(*Taken from Casella and Berger Theorem 5.5.2*)

Theorem A.1.4 Suppose $X_1, X_2, ...$ converges in probability to a random variable X and that h is a continuous function. Then, $h(X_1), h(X_2), ...$ converges in probability to $h(X)$.

(*Taken from Casella and Berger Theorem 5.5.4*)

Theorem A.1.5 Multivariate Central Limit Theorem Let $X_j = X_{j1}, ..., X_{jn}$, $j = 1, ..., p$ be *iid* random vectors with p dimensional mean vector $\mu = (\mu_1, .., \mu_p)^T$ and variance-covariance matrix, Σ, then,

$$[\sqrt{n}(\bar{X}_{n1} - \mu_1), ..., \sqrt{n}(\bar{X}_{np} - \mu_p)] \xrightarrow{D} N_p(0, \Sigma), \tag{A.2}$$

where $\bar{X}_{nr} = \frac{1}{n}\sum_{j=1}^{n} X_{jr}$.

(*Taken from Lehmann and Casella, Theorem 8.21*)

Theorem A.1.6: The Delta Method Suppose that $[\sqrt{n}(X_{n1}-\mu_1), ..., \sqrt{n}(X_{np}-\mu_p)] \xrightarrow{D} N_p(0, \Sigma)$ and let $h_1, ..., h_p$ be p real-valued functions of $\gamma = (\gamma_1, ..., \gamma_p)^T$ that are defined and continuously differentiable in a neighbourhood of the parameter point γ. Next, define the nonsingular matrix, $\nabla = \left\{\frac{\partial h_i}{\partial \gamma_j}\right\}_{i,j}$. Then,

$$[\sqrt{n}(h_1(X_{n1}) - h_1(\mu_1)), ..., \sqrt{n}(h_p(X_{np}) - h_p(\mu_p))] \xrightarrow{D} N_p(0, \nabla\Sigma\nabla^T), \tag{A.3}$$

(*Taken from Lehmann and Casella, Theorem 8.22*)

Theorem A.1.7 Consider $\eta_i = h_i(\gamma)$ such that h is a one-to-one transformation and $h_i^{-1}(\eta) = \gamma_i$ exists. Then, the Fisher Information for the new parametrization will be,

$$\mathcal{I}^*(\gamma) = J\mathcal{I}(\eta)J^T \tag{A.4}$$

where J is the Jacobian matrix with elements $\frac{\partial \eta_i}{\partial \gamma_j}$.

(*Adapted from Lehmann Elements of Large Sample Theory*)

Definition A.1.1: Schur Complement Let A be an $m \times m$ matrix, B be an $m \times n$ matrix, C be an $n \times m$ matrix and D be an $n \times n$ matrix. If D is nonsingular, and we define the partitioned matrix T to be

$$T = \left[\begin{array}{c:c} A & B \\ \hdashline C & D \end{array}\right], \tag{A.5}$$

then the Schur complement of A in T is $Q = A - BD^{-1}C$.

Theorem A.1.8: Positive definiteness of a Schur Complement If T is positive definite, then the Schur complement Q is also positive definite.

(*Taken from Harville 1997 Theorem 14.8.4*)

A.2 Multi-Sample Results

Definition A.2.1 Multi-Sample Data Let $X_{s1}, ..., X_{sn_s}$ be the s^{th} independently drawn sample $(s = 1, ..., S)$ from sample space $\mathcal{X}_s$ of *iid* random variables with density f_s, cumulative distribution F_s, and common p-dimensional parameter vector, $\gamma \in \Gamma$. Next, assume that the sample proportions converge to weights, ω_s, such that $0 < \omega_s < 1$, $\sum_{s=1}^{S} \omega_s = 1$ for all s and $\left(\frac{n1}{n}, ..., \frac{n_s}{n}\right) \longrightarrow (\omega_1, ..., \omega_s)$ where $\sum_{i=1}^{n} n_i = n$. Then, $X_{s1}, ..., X_{sn_s}$ can be defined as multi-sample data. Finally, the overall combined likelihood for the data will be $L(\gamma|\mathbf{x}) = \prod_{s=1}^{S} \prod_{i=1}^{n_s} f_s(x_{si}; \gamma)$ and the log likelihood will be $l(\gamma|\mathbf{x}) = \sum_{s=1}^{S} \sum_{i=1}^{n_s} \log f_s(x_{si}; \gamma)$.

(*Adapted from Hirose Section 6*)

Theorem A.2.1 Multi-Sample Weak Law of Large Numbers Let $X_{s1}, ..., X_{sn_s}$ be multi-sample data, such that $Var(X_{si}) < \infty$ and let h_s be a continuous function. Next define $E_s(h_s(X)) = \int h_s(x) f_s(x;\gamma)dx$, where f_s represents the density of X_s. Then,

$$\frac{1}{n}\sum_{s=1}^{S}\sum_{i=1}^{n_s} h_s(X_{si}) \xrightarrow{p} \sum_{s=1}^{S}\omega_s E_s(h_s(X)) \tag{A.6}$$

Proof

$$\frac{1}{n}\sum_{s=1}^{S}\sum_{i=1}^{n_s} h_s(X_{si}) = \sum_{s=1}^{S}\frac{n_s}{n}\frac{1}{n_s}\sum_{i=1}^{n_s} h_s(X_{si}) \xrightarrow{p} \sum_{s=1}^{S}\omega_s E(h_s(X)), \tag{A.7}$$

since,

$$\frac{1}{n_s}\sum_{i=1}^{n_s} h_s(X_{si}) \xrightarrow{p} E(h_s(X)), \tag{A.8}$$

and $\left(\frac{n_1}{n}, ..., \frac{n_s}{n}\right) \longrightarrow (\omega_1, ..., \omega_s)$, by Theorem A.1.2, Theorem A.1.3 (WLLN) and Theorem A.1.4.

(*Adapted from Hirose Lemma 6.1*)

A.2.2 Multi-sample Central Limit Theorem Let $X_{s1}, ..., X_{sn_s}$ be multi-sample data with $E_s(h_s(X)) = 0$ and $Var(X_{si}) < \infty$ for all s and let h_s be a continuous function, then,

$$\frac{1}{\sqrt{n}}\sum_{s=1}^{S}\sum_{i=1}^{n_s} h_s(X_{si}) \xrightarrow{D} N(0, \sum_{s=1}^{S}\omega_s\Sigma_s), \tag{A.9}$$

where $\Sigma_s = E_s[h_s(X)h_s(X)^T]$.

Proof By the single sample CLT, independence between samples, Slutsky's theorem and the assumption $\left(\frac{n1}{n}, ..., \frac{n_s}{n}\right) \longrightarrow (\omega_1, ..., \omega_s)$,

$$\frac{1}{\sqrt{n}} \sum_{s=1}^{S} \sum_{i=1}^{n_s} h_s(X_{si}) = \sum_{s=1}^{S} \sqrt{\frac{n_s}{n}} \frac{1}{\sqrt{n_s}} \sum_{i=1}^{n_s} h_s(X_{si}) \xrightarrow{D} \sum_{s=1}^{S} \sqrt{\omega_s} Y_s, \tag{A.10}$$

since

$$\frac{1}{\sqrt{n_s}} \sum_{i=1}^{n_s} h_s(X_{si}) \xrightarrow{D} Y_s,$$

where $Y_s \sim N(0, \Sigma_s)$, since $E_s(h_s(X)) = 0$ for all s.

Finally, $\sum_{s=1}^{S} \sqrt{\omega_s} Y_s \sim N(0, \sum_{s=1}^{S} \omega_s \Sigma_s)$ due to the independence of the s samples.

(*Adapted from Hirose Lemma 6.1*)

Theorem A.2.3 Let $X_{s1}, ..., X_{sn_s}$ be multi-sample data that satisfies the following regularity conditions (adapted from Lehmann, Casella and Hirose),

i) There exists an open subset, Γ_0 of parameter space Γ containing the true parameter point, γ_0.

ii) For every $x \in \mathcal{X}_s$, the density, $f_s(x|\gamma)$, admits all third derivatives, $\frac{\partial^3}{\partial\gamma_j \partial\gamma_k \partial\gamma_l} f_s(x|\gamma)$, for all $\gamma \in \Gamma$, which are bounded by functions with finite expectations.

iii) There exists functions M_{sjkl} with finite expectation such that

$$\left| \frac{\partial^3}{\partial\gamma_j \partial\gamma_k \partial\gamma_l} f_s(x|\gamma) \right| \leq M_{sjkl}(x)$$

for all $\gamma \in \Gamma$, j, k, l and $s = 1, ..., S$.

v) $f_s(\cdot;\gamma)$ satisfies,

$$\begin{aligned} &a) \quad E[U_s(\gamma)] = 0 \\ &b) \quad \mathcal{I}_s(\gamma) = E[U_s(\gamma)U_s(\gamma)^T] = E[-\frac{\partial^2 l_s(\gamma)}{\partial\gamma\partial\gamma^T}] \end{aligned} \tag{A.11}$$

where $U_s(\gamma)$ is the score vector for the s^{th} sample with elements $\frac{\partial}{\partial\gamma_j} l_s(\gamma|x)$, $j = 1, ..., p$, and $\mathcal{I}_s(\gamma)$ is the $p \times p$ positive definite Fisher's information matrix for the s^{th} sample, $s = 1, ..., S$.

Then, as $n \longrightarrow \infty$, $\sum_s U_s(\gamma) = 0$ has the solution, $\hat{\gamma}_n$ such that,

$$\begin{aligned} &i) \quad \hat{\gamma}_n \xrightarrow{p} \gamma_0 \\ &ii) \quad \sqrt{n}(\hat{\gamma}_n - \gamma_0) \xrightarrow{D} N\left(0, \left[\sum_{s=1}^{s} \omega_s \mathcal{I}_s(\gamma_0)\right]^{-1}\right) \end{aligned} \tag{A.12}$$

with probability 1.

Proof of i) To demonstrate $\hat{\gamma}_n \xrightarrow{p} \gamma_0$, the single sample proof will hold, except for demonstrating that the s-sample likelihood is maximized at γ_0. Thus, we will demonstrate this first, followed by a reproduction of the rest of the proof as demonstrated in Hogg & Craig, Ch. 6.

Theorem A.2.3.1 The s-sample likelihood, $L(\gamma|x) = \prod_{s=1}^{S}\prod_{i=1}^{n_s} f_s(x_{si};\gamma)$ is maximized at γ_0.

Proof Under regularity condition (i) and Definition A.2.1, we will demonstrate $\lim\limits_{n\longrightarrow\infty} P[L(\gamma_0|x) > L(\gamma|x)] = 1$ for all $\gamma \neq \gamma_0$.

Assume, $L(\gamma_0|X) > L(\gamma|X)$ for all $\gamma \neq \gamma_0$, then,

$$\frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{n_s} \log f_s(X_{si};\gamma_0) > \frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{n_s} \log f_s(X_{si};\gamma),$$

which implies,

$$\frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{n_s}\log\left[\frac{f_s(X_{si};\gamma)}{f_s(X_{si};\gamma_0)}\right] < 0.$$

By Theorem A.2.1,

$$\frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{n_s}\log\left[\frac{f_s(X_{si};\gamma)}{f_s(X_{si};\gamma_0)}\right] \xrightarrow{p} \sum_{s=1}^{S}\omega_s E_{s,\gamma_0}\log\left[\frac{f_s(X;\gamma)}{f_s(X;\gamma_0)}\right],$$

and by Jensen's inequality,

$$\sum_{s=1}^{S}\omega_s E_{s,\gamma_0}\log\left[\frac{f_s(X;\gamma)}{f_s(X;\gamma_0)}\right] < \sum_{s=1}^{S}\omega_s \log E_{s,\gamma_0}\left[\frac{f_s(X;\gamma)}{f_s(X;\gamma_0)}\right].$$

Finally, $\log E_{s,\gamma_0}\left[\frac{f_s(X;\gamma)}{f_s(X;\gamma_0)}\right] = 0$ for all $s = 1, ..., S$ since,

$$E_{s,\gamma_0}\left[\frac{f_s(X;\gamma)}{f_s(X;\gamma_0)}\right] = \int \frac{f_s(x;\gamma)}{f_s(x;\gamma_0)} f_s(x;\gamma_0)dx = 1.$$

Proof of (i), con't From regularity condition (i), $(\gamma_0 - a, \gamma_0 + a) \subset \Gamma$ for $a > 0$. Let,

$$S_n = \{x : l(\gamma_0; x) > l(\gamma_0 - a; x)\} \cap \{x : l(\gamma_0; x) > l(\gamma_0 + a; x)\},$$

where $l()$ denotes the log-likelihood under a multi-sample as defined in Definition A.2.1, then, by Theorem A.2.3.1, $P(S_n) \longrightarrow 1$.

$$S_n \subset \{x : |\hat{\gamma}_n - \gamma_0| < a\} \cap \{x : l'(\hat{\gamma}_n|x) = 0\},$$

$$1 = \lim_{n\longrightarrow\infty} P(S_n) \leq \limsup_{n\longrightarrow\infty} P[\{x : |\hat{\gamma}_n - \gamma_0| < a\} \cap \{x : l'(\hat{\gamma}_n|x) = 0\}] \leq 1.$$

which implies that $P(|\hat{\gamma}_n - \gamma_0| < a) \longrightarrow 1$ where $\hat{\gamma}_n$ is the solution to $l'(\gamma|x) = 0$. *(Adapted from Hogg & Craig Theorem 6.1.3 p. 316)*

Proof of ii) Using a Taylor series expansion about γ_0,

$$\sum_{s=1}^{S} U_s(\hat{\gamma}_n) = \sum_{s=1}^{S} U_s(\gamma_0) + (\hat{\gamma}_n - \gamma_0)\sum_{s=1}^{S}\frac{\partial}{\partial\gamma}U_s(\gamma)\Big|_{\gamma=\gamma_0} + ...$$

which implies,

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0) \approx \frac{1}{\sqrt{n}} \sum_{s=1}^{S} U_s(\gamma_0) \left[-\frac{1}{n} \sum_{s=1}^{S} \frac{\partial}{\partial \gamma} U_s(\gamma) \Big|_{\gamma=\gamma_0} \right]^{-1}$$

By Theorem A.2.1 and the convergence of the sample proportions,

$$\left[-\frac{1}{n} \sum_{s=1}^{S} \frac{\partial}{\partial \gamma} U_s(\gamma) \Big|_{\gamma=\gamma_0} \right] \xrightarrow{p} -\sum_{s=1}^{S} \omega_s E_s \left[\frac{\partial}{\partial \gamma} U_s(\gamma) \Big|_{\gamma=\gamma_0} \right] = -\sum_{s=1}^{S} \omega_s \mathcal{I}_s(\gamma_0).$$

Next, by Theorem A.2.2,

$$\frac{1}{\sqrt{n}} \sum_{s=1}^{S} U_s(\gamma_0) \xrightarrow{D} \sum_{s=1}^{S} \sqrt{\omega_s} Y_s \tag{A.13}$$

such that $Y_s \sim N(0, \Sigma_s)$ where $\Sigma_s = E[U_s(\gamma_0)U_s(\gamma_0)^T] = \mathcal{I}_s(\gamma_0)$ which implies $\sum_{s=1}^{S} \sqrt{\omega_s} Y_s \sim N(0, \sum_{s=1}^{S} \omega_s \mathcal{I}_s(\gamma_0))$.

Therefore, by Slutsky's,

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0) \xrightarrow{D} \left(\sum_{s=1}^{S} \sqrt{\omega_s} Y_s \right) \left(-\sum_{s=1}^{S} \omega_s \mathcal{I}_s(\gamma_0) \right)^{-1} \sim N \left(0, \left[\sum_{s=1}^{S} \omega_s \mathcal{I}_s(\gamma_0) \right]^{-1} \right). \tag{A.14}$$

(*Adapted from Hirose Theorem 6.1*)

Theorem A.2.4 Under a multi-sample framework and the conditions of Theorem A.2.3, the observed information matrix evaluated at $\hat{\gamma}_n$ is a consistent estimator of $\sum_{s=1}^{S} \omega_s \mathcal{I}_s(\gamma_0)$.

Proof Let $I(\gamma) = -\frac{1}{n} \frac{\partial l(\gamma|\mathbf{x})}{\partial \gamma \partial \gamma^T}$. Then, by Slutsky's Theorem and Theorem 3.2.1,

$$I(\gamma) = \sum_{s=1}^{S} \frac{n_s}{n} \frac{1}{n_s} \sum_{i=1}^{n_s} \left[-\frac{\partial l_s(\gamma|\mathbf{x}_s)}{\partial \gamma \partial \gamma^T} \right] \xrightarrow{p} \sum_{s=1}^{S} \omega_s \mathcal{I}_s(\gamma). \tag{A.15}$$

Hence, $I(\gamma)|_{\gamma=\hat{\gamma}_n}$ is a consistent estimator of $\sum_{s=1}^{S} \omega_s \mathcal{I}_s(\gamma_0)$ since $\hat{\gamma}_n \xrightarrow{p} \gamma_0$ by assumption.

Theorem A.2.5: The multi-parameter maximum likelihood estimate, $\hat{\gamma}$, derived from a s-sample framework has minimum variance

Proof Let $\hat{\zeta} = (\hat{\zeta}_1, ..., \hat{\zeta}_p)^T$ be any p-dimensional s-sample estimator satisfying the regularity conditions on Theorem A.2.3. Recall, from the proof of Theorem A.2.3, the total score, $\sum_{s=1}^{S} U_s(\zeta_0)$, has asymptotic variance-covariance matrix $n\sum_{s=1}^{S} \omega_s \mathcal{I}_s(\zeta_0)$.

Then

$$\begin{aligned} Cov\left[\begin{array}{c} \hat{\zeta} \\ \hline \sum_{s=1}^{S} U_s(\zeta_0) \end{array}\right] &= \left[\begin{array}{c|c} Cov(\hat{\zeta}) & E_\zeta(\hat{\zeta}[\sum_{s=1}^{S} U_s(\zeta_0)]^T) \\ \hline E_\zeta([\sum_{s=1}^{S} U_s(\zeta_0)]\hat{\zeta}^T) & Cov(\sum_{s=1}^{S} U_s(\zeta_0)) \end{array}\right] \qquad \text{(A.16)} \\ &= \left[\begin{array}{c|c} Cov(\hat{\zeta}) & SI_p \\ \hline SI_p & n\sum_{s=1}^{S} \omega_s \mathcal{I}_s(\zeta_0) \end{array}\right], \end{aligned}$$

where I_p is the p-dimensional identity matrix and since,

$$\begin{aligned} E_\zeta(\hat{\zeta}[\sum_{s=1}^{S} U_s(\zeta_0)]^T) &= E_\zeta(\hat{\zeta}[\sum_{s=1}^{S} \frac{\partial}{\partial \zeta} \log f_s(X|\zeta)]^T) \qquad \text{(A.17)} \\ &= E_\zeta(\hat{\zeta}[\sum_{s=1}^{S} \frac{\partial f_s(X|\zeta)}{\partial \zeta} \frac{1}{f_s(\mathbf{x}_s|\zeta)}]^T) \\ &= \sum_{s=1}^{S} \frac{\partial}{\partial \zeta} E_\zeta[\hat{\zeta}] \\ &= S, \end{aligned}$$

since $\hat{\zeta}$ is unbiased for ζ.

Finally,

$$Cov\left[\begin{array}{c} \hat{\zeta} \\ \hline \sum_{s=1}^{S} U_s(\zeta_0) \end{array}\right] \succ 0, \qquad \text{(A.18)}$$

since it is a covariance matrix implying that its Schur complement is positive definite as well by Theorem A.1.8.

Therefore,

$$Cov(\hat{\zeta}) - S^2[n\sum_{s=1}^{S}\omega_s\mathcal{I}_s(\zeta_0)]^{-1} \succ 0, \tag{A.19}$$

which implies

$$\begin{aligned} Cov(\hat{\zeta}) &\succ S^2[n\sum_{s=1}^{S}\omega_s\mathcal{I}_s(\zeta_0)]^{-1} \\ &\succ [n\sum_{s=1}^{S}\omega_s\mathcal{I}_s(\zeta_0)]^{-1}, \end{aligned} \tag{A.20}$$

since S is a positive integer.

(Adapted from Greer et al. 2009)

Appendix B

Derivation of Results for Chapter 5

In this appendix, we will outline in detail the justification for some of the results described in Chapter 5. We develop these results allowing for the misclassification probabilities to depend on time; however, we will use an asterisk (eg. $\overset{*}{=}$) to signify that the assumptions, $\theta(t) = \theta$, $\phi(t) = \phi$ and $\xi(t) = \xi$, have been applied.

B.1 Univariate Model, $L^{(1)}$

First, we will redefine the misclassification probabilities, keeping in mind the assumption that misclassification can only occur if a true AE has occurred,

$$P[\tilde{T} \in I_{dt} | T \in I_{du}] = \begin{cases} 0 & u < t \\ \theta(t) & u = t \\ 0 & u > t \end{cases} \tag{B.1}$$

$$P[\tilde{T} > t | T \in I_{du}] = \begin{cases} 0 & u < t \\ 1 - \theta(t) & u = t \\ 1 & u > t \end{cases} \tag{B.2}$$

at time t. Note that the assumption that a true failure must occur for misclassification to be possible makes consideration of times $u : u < t$ irrelevant, hence, probability zero.

i) $P[\tilde{T} \in I_{dt}] \overset{*}{=} \theta f(t)$

$$\begin{aligned}
P[\tilde{T} \in I_{dt}] &= \int_0^\infty P(\tilde{T} \in I_{dt}|T \in I_{du}) f(u) du \\
&= \int_0^t P[\tilde{T} \in I_{dt}|T \in I_{du}] f(u) du + \int_{t^+}^\infty P[\tilde{T} \in I_{dt}|T \in I_{du}] f(u) du \\
&= P[\tilde{T} \in I_{dt}|T \in I_{dt}] f(t) dt + 0 \\
&= \theta(t) f(t) dt \\
&\overset{*}{=} \theta f(t) dt
\end{aligned}$$

by equation (B.1).

ii) $P[\tilde{T} > t] \overset{*}{=} 1 - \theta(1 - S(t))$

$$\begin{aligned}
P[\tilde{T} > t] &= \int_0^t P[\tilde{T} > t|T \in I_{du}] f(u) du + \int_{t^+}^\infty P[\tilde{T} > t|T \in I_{du}] f(u) du \\
&= \int_0^t (1 - P[\tilde{T} \le u|T \in I_{du}]) f(u) du + \int_{t^+}^\infty f(u) du \\
&= \int_0^t f(u) du - \int_0^t \int_0^t P[\tilde{T} \in I_{dv}|T \in I_{du}] f(u) du dv + S(t) \\
&= 1 - \int_{0<u=v<t} P[\tilde{T} \in I_{du}|T \in I_{du}] f(u) du \\
&= 1 - \int_0^t \theta(u) f(u) du \\
&\overset{*}{=} 1 - \theta(1 - S(t))
\end{aligned}$$

by (B.1), $P[\tilde{T} \in I_{dv}|T \in I_{du}] = 0$ unless $u = v$ for $u, v < t$.

iii) $P[T \leq t|\tilde{T} > t] \overset{*}{=} \frac{(1-\theta)(1-S(t))}{1-\theta(1-S(t))}$

$$\begin{aligned}
P[T \leq t|\tilde{T} > t] &= \frac{(1 - P[\tilde{T} \leq t|T \leq t])P(T \leq t)}{P[\tilde{T} > t]} \\
&= \frac{(1 - S(t)) - \int_0^t P[\tilde{T} \in I_{du}|T \in I_{du}]f(u)du}{P[\tilde{T} > t]} \\
&= \frac{(1 - S(t)) - \int_0^t \theta(u)f(u)du}{1 - \int_0^t \theta(u)f(u)du} \\
&\overset{*}{=} \frac{(1 - \theta)(1 - S(t))}{1 - \theta(1 - S(t))}
\end{aligned}$$

iv) $P[T > t|\tilde{T} > t] \overset{*}{=} \frac{S(t)}{1-\theta(1-S(t))}$

$$\begin{aligned}
P[T > t|\tilde{T} > t] &= \frac{P[\tilde{T} > t|T > t]P[T > t]}{P(\tilde{T} > t)} \\
&= \frac{S(t)}{1 - \int_0^t \theta(u)f(u)du} \\
&\overset{*}{=} \frac{S(t)}{(1 - \theta(1 - S(t))}
\end{aligned}$$

B.2 Competing Risks Model, $L^{(2)}$

For this model, we will restate the definition of the time-to-event misclassification probabilities to incorporate the C-S hazards,

$$P[\tilde{T} \in I_{dt} | T \in I_{du}, \tilde{J} = j] = \begin{cases} 0 & u < t \\ \phi(t) & u = t \\ 0 & u > t \end{cases} \quad \text{(B.3)} \qquad P[\tilde{T} > t | T \in I_{du}, \tilde{J} = j] = \begin{cases} 0 & u < t \\ 1 - \phi(t) & u = t \\ 1 & u > t \end{cases}, \quad \text{(B.4)}$$

for $j = 1, 2$. Next, note that,

$$\begin{aligned} P[\tilde{J} = 1 | T \in I_{dt}] &= P[\tilde{J} = 1 | J = 1, T \in I_{dt}] P[J = 1 | T \in I_{dt}] \qquad \text{(B.5)} \\ &\quad + P[\tilde{J} = 1 | J = 2, T \in I_{dt}] P[J = 2 | T \in I_{dt}] \\ &= \xi_1(t) \frac{\lambda_1(t)}{\lambda(t)} + (1 - \xi_2(t)) \frac{\lambda_2(t)}{\lambda(t)} \\ &\overset{*}{=} \xi_1 \frac{\lambda_1(t)}{\lambda(t)} + (1 - \xi_2) \frac{\lambda_2(t)}{\lambda(t)} \\ P[\tilde{J} = 2 | T \in I_{dt}] &\overset{*}{=} (1 - \xi_1) \frac{\lambda_1(t)}{\lambda(t)} + \xi_2 \frac{\lambda_2(t)}{\lambda(t)} \end{aligned}$$

i) $P[\tilde{T} \in I_{dt}, \tilde{J} = 1] \overset{*}{=} \phi S(t)[\xi_1 \lambda_1(t) + (1 - \xi_2)\lambda_2(t)]dt$ and $P[\tilde{T} \in I_{dt}, \tilde{J} = 2] \overset{*}{=} \phi S(t)[(1 - \xi_1)\lambda_1(t) + \xi_2 \lambda_2(t)]dt$

$$\begin{aligned} P[\tilde{T} \in I_{dt}, \tilde{J} = j] &= \int_0^\infty P[\tilde{T} \in I_{dt} | T \in I_{dt}, \tilde{J} = j] f_j(u) du \\ &= P[\tilde{T} \in I_{dt} | T \in I_{dt}, \tilde{J} = j] P[T \in I_{dt}, \tilde{J} = j] + 0 \\ &= P[\tilde{T} \in I_{dt} | T \in I_{dt}, \tilde{J} = j] P[\tilde{J} = j | T \in I_{dt}] P[T \in I_{dt}] \\ &= \phi(t) P[\tilde{J} = j | T \in I_{dt}] f(t) dt \end{aligned}$$

thus,

$$P[\tilde{T} \in I_{dt}, \tilde{J} = 1] \overset{*}{=} \phi S(t)[\xi_1 \lambda_1(t) + (1 - \xi_2)\lambda_2(t)]dt,$$

and

$$P[\tilde{T} \in I_{dt}, \tilde{J} = 2] \quad \overset{*}{=} \quad \phi S(t)[(1 - \xi_1)\lambda_1(t) + \xi_2 \lambda_2(t)]dt.$$

ii) $P[J = 1|\tilde{J} = 1, T \in I_{dt}] \overset{*}{=} \frac{\xi_1 \lambda_1(t)}{\xi_1 \lambda_1(t) + (1-\xi_2)\lambda_2(t)}$

$P[J = 2|\tilde{J} = 1, T \in I_{dt}] \overset{*}{=} \frac{(1-\xi_2)\lambda_2(t)}{\xi_1 \lambda_1(t) + (1-\xi_2)\lambda_2(t)}$

$P[J = 1|\tilde{J} = 2, T \in I_{dt}] \overset{*}{=} \frac{(1-\xi_1)\lambda_1(t)}{(1-\xi_1)\lambda_1(t) + \xi_2 \lambda_2(t)}$

$P[J = 2|\tilde{J} = 2, T \in I_{dt}] \overset{*}{=} \frac{\xi_2 \lambda_2(t)}{(1-\xi_1)\lambda_1(t) + \xi_2 \lambda_2(t)}$

These all follow from,

$$P[J = j|\tilde{J} = k, T \in I_{dt}] = \frac{P[\tilde{J} = k|J = j, T \in I_{dt}]P[J = j|T \in I_{dt}]}{\sum_{j=1,2} P[\tilde{J} = k|J = j, T \in I_{dt}]P[J = j|T \in I_{dt}]}.$$

B.3 Competing Risks Model, $L^{(2b)}$

i) $P[T \in I_{dt}, \tilde{J} = 1] \overset{*}{=} S(t)[\xi_1 \lambda_1(t) + (1 - \xi_2)\lambda_2(t)]dt$

$P[T \in I_{dt}, \tilde{J} = 2] \overset{*}{=} S(t)[(1 - \xi_1)\lambda_1(t) + \xi_2 \lambda_2(t)]dt,$

Since,

$$\begin{aligned} P[T \in I_{dt}, \tilde{J} = j] &= P[\tilde{J} = j|T \in I_{dt}]P[T \in I_{dt}] \\ &= P[\tilde{J} = j|T \in I_{dt}]f(t)dt \end{aligned}$$

where $P[\tilde{J} = j|T \in I_{dt}]$ is equation (B.5).

Appendix C

R code

C.1 Chapter 2 Code

Data Generation

```
datagenCont<-function(N1,N0,P11,P10,Th11=1,Th01=1, Th10=0, Th00=0) {
#Generate dataset with misclassification(d,a,\tilde{a})
P01<-1-P11 #pi 0|1
P00<-1-P10 #pi 0|0

N11<-rbinom(1,N1,P11)
N01<-rbinom(1,N0,P10)
N10<-N1-N11
N00<-N0-N01

trueclass11<-rbinom(1,N11,Th11)
trueclass10<-rbinom(1,N10,1-Th10)
trueclass01<-rbinom(1,N01,Th01)
trueclass00<-rbinom(1,N00,1-Th00)

misclass11<-N11-trueclass11
misclass10<-N10-trueclass10
misclass01<-N01-trueclass01
misclass00<-N00-trueclass00

dataset<-rbind(matrix(rep(c(1,1,1),trueclass11),nr=trueclass11,nc=3,byrow=TRUE),
matrix(rep(c(1,0,1),misclass10),nr=misclass10,nc=3,byrow=TRUE),
matrix(rep(c(0,1,1),trueclass01),nr=trueclass01,nc=3,byrow=TRUE),
matrix(rep(c(0,0,1),misclass00),nr=misclass00,nc=3,byrow=TRUE),
matrix(rep(c(1,1,0),misclass11),nr=misclass11,nc=3,byrow=TRUE),
matrix(rep(c(1,0,0),trueclass10),nr=trueclass10,nc=3,byrow=TRUE),
matrix(rep(c(0,1,0),misclass01),nr=misclass01,nc=3,byrow=TRUE),
matrix(rep(c(0,0,0),trueclass00),nr=trueclass00,nc=3,byrow=TRUE))

n11<-trueclass11+misclass10
n10<-trueclass10+misclass11
n01<-trueclass01+misclass00
n00<-trueclass00+misclass01

list(N=c(N11,N10,N01,N00),n=c(n11,n10,n01,n00),D=dataset)
}
```

2S Approach: Validation Sampling Algorithm

```
Val2Sample<-function(M,dataset) {
```

```
y11 = 0; y10 = 0; y01 = 0; y00 = 0;
errorCheck<-0
while(y11 == 0 || y10 == 0 || y01 == 0 || y00 == 0 || y11 == m11 ||
  y10 == m10 || y01 == m01 || y00 == m00) {

valInd<-sample(1:length(dataset[,1]),M)
valSample<-rbind(dataset[valInd,])

m11<-length(which(valSample[,1]==1 & valSample[,2]==1 & valSample[,3]==1))+
length(which(valSample[,1]==1 & valSample[,2]==0 & valSample[,3]==1))

m01<-length(which(valSample[,1]==0 & valSample[,2]==1 & valSample[,3]==1))+
length(which(valSample[,1]==0 & valSample[,2]==0 & valSample[,3]==1))

m10<-length(which(valSample[,1]==1 & valSample[,2]==0 & valSample[,3]==0))+
length(which(valSample[,1]==1 & valSample[,2]==1 & valSample[,3]==0))

m00<-length(which(valSample[,1]==0 & valSample[,2]==1 & valSample[,3]==0))+
length(which(valSample[,1]==0 & valSample[,2]==0 & valSample[,3]==0))

y11<-length(which(valSample[,1]==1 & valSample[,2]==1 & valSample[,3]==1))
y01<-length(which(valSample[,1]==0 & valSample[,2]==1 & valSample[,3]==1))
y10<-length(which(valSample[,1]==1 & valSample[,2]==1 & valSample[,3]==0))
y00<-length(which(valSample[,1]==0 & valSample[,2]==1 & valSample[,3]==0))
errorCheck<-errorCheck+1
 errorCheck<-errorCheck+1
 if(errorCheck>1000) {print("Error stopped after 100 attempts")
break}
  }
list(y=c(y11,y10,y01,y00),m=c(m11,m10,m01,m00))
}
```

3S Approach: Validation Sampling Algorithm

```
Val3Sample<-function(M1,M0,dataset) {

n1 = sum(dataset[,3]); # n_atilde =1
n0 = length(dataset[,1])-n1; # n_atilde =0
y11 = 0; y10 = 0; y01 = 0; y00 = 0;
errorCheck<-0

while(y11 == 0 || y10 == 0 || y01 == 0 || y00 == 0 || y11 == m11 ||
  y10 == m10 || y01 == m01 || y00 == m00) {


valInd1<-sample(1:n1,M1)
valInd0<-sample((n1+1):(n1+n0),M0)

valSample1<-rbind(dataset[valInd1,])
valSample0<-rbind(dataset[valInd0,])

m11<-length(which(valSample1[,1]==1 & valSample1[,2]==1 & valSample1[,3]==1))+
length(which(valSample1[,1]==1 & valSample1[,2]==0 & valSample1[,3]==1))
m01<-length(which(valSample1[,1]==0 & valSample1[,2]==1 & valSample1[,3]==1))+
length(which(valSample1[,1]==0 & valSample1[,2]==0 & valSample1[,3]==1))

m10<-length(which(valSample0[,1]==1 & valSample0[,2]==0 & valSample0[,3]==0))+
length(which(valSample0[,1]==1 & valSample0[,2]==1 & valSample0[,3]==0))
m00<-length(which(valSample0[,1]==0 & valSample0[,2]==1 & valSample0[,3]==0))+
length(which(valSample0[,1]==0 & valSample0[,2]==0 & valSample0[,3]==0))

y11<- length(which(valSample1[,1]==1 & valSample1[,2]==1 & valSample1[,3]==1))
y01<-length(which(valSample1[,1]==0 & valSample1[,2]==1 & valSample1[,3]==1))
y10<-length(which(valSample0[,1]==1 & valSample0[,2]==1 & valSample0[,3]==0))
y00<-length(which(valSample0[,1]==0 & valSample0[,2]==1 & valSample0[,3]==0))
 errorCheck<-errorCheck+1
 if(errorCheck>1000) {print("Error stopped after 100 attempts")
break}
  }
list(y=c(y11,y10,y01,y00),m=c(m11,m10,m01,m00))
}
```

Information Matrix - Original Sample

```
I_O_Cont<-function(n,param) {
n11<-n[1]
n10<-n[2]
n01<-n[3]
n00<-n[4]

P11_mle<-param[1]
P10_mle<-param[2]

Phi11_mle<-param[3]
Phi10_mle<-param[4]
Phi01_mle<-param[5]
Phi00_mle<-param[6]
IS11 <- 0;
IS22 <- 0;
IS33 <- n11/(Phi11_mle + Phi01_mle)^2 + n10/(1 - Phi11_mle - Phi01_mle)^2;
IS44 <- n01/(Phi10_mle + Phi00_mle)^2 + n00/(1 - Phi10_mle - Phi00_mle)^2;
IS55 <- IS33;
IS66 <- IS44;

IS12 <- 0;
IS13 <- 0;
IS14 <- 0;
IS15 <- 0;
IS16 <- 0;
IS23 <- 0;
IS24 <- 0;
IS25 <- 0;
IS26 <- 0;
IS34 <- 0;
IS35 <- IS33;
IS36 <- 0;
IS45 <- 0;
IS46 <- IS44;
IS56 <- 0;

IS<-rbind(c(IS11, IS12, IS13, IS14, IS15, IS16),c(IS12, IS22, IS23, IS24, IS25, IS26),
c(IS13, IS23, IS33, IS34, IS35, IS36),c(IS14, IS24, IS34, IS44, IS45, IS46),
c(IS15, IS25, IS35, IS45, IS55, IS56),c(IS16, IS26, IS36, IS46, IS56, IS66))

return(IS)
}
```

Information Matrix - 2S Approach Validation Sample

```
I_V_Cont<-function(m,y,param) {
m2.11<-m[1]
m2.10<-m[2]
m2.01<-m[3]
m2.00<-m[4]

y2.11<-y[1]
y2.10<-y[2]
y2.01<-y[3]
y2.00<-y[4]

P11_mle<-param[1]
P10_mle<-param[2]

Phi11_mle<-param[3]
Phi10_mle<-param[4]
Phi01_mle<-param[5]
Phi00_mle<-param[6]

IV11 <- (m2.10 - y2.10)/(1 - P11_mle - Phi01_mle)^2 + y2.10/(P11_mle - Phi11_mle)^2;
IV22 <- (m2.00 - y2.00)/(1 - P10_mle - Phi00_mle)^2 + y2.00/(P10_mle - Phi10_mle)^2;
IV33 <- y2.10/(P11_mle - Phi11_mle)^2 +  y2.11/Phi11_mle^2;
IV44 <- y2.00/(P10_mle - Phi10_mle)^2 + y2.01/Phi10_mle^2 ;

IV55 <-(m2.10 - y2.10)/(1 - P11_mle - Phi01_mle)^2 + (m2.11 - y2.11)/Phi01_mle^2 ;
IV66 <- (m2.00 - y2.00)/(1 - P10_mle - Phi00_mle)^2 + (m2.01 - y2.01)/Phi00_mle^2;

IV12 <- 0;
IV13 <- -(y2.10/(P11_mle - Phi11_mle)^2);
IV14 <- 0;
IV15 <- (m2.10 - y2.10)/(1 - P11_mle - Phi01_mle)^2;
IV16 <- 0;
IV23 <- 0;
IV24 <- -(y2.00/(P10_mle - Phi10_mle)^2);
IV25 <- 0;
IV26 <- (m2.00 - y2.00)/(1 - P10_mle - Phi00_mle)^2;
IV34 <- 0;
IV35 <- 0;
IV36 <- 0;
IV45 <- 0;
IV46 <- 0;
IV56 <- 0;

IV<-rbind(c(IV11, IV12, IV13, IV14, IV15, IV16),c(IV12, IV22, IV23, IV24, IV25, IV26),
c(IV13, IV23, IV33, IV34, IV35, IV36),c(IV14, IV24, IV34, IV44, IV45, IV46),
c(IV15, IV25, IV35, IV45, IV55, IV56),c(IV16, IV26, IV36, IV46, IV56, IV66))
return(IV)
}
```

Information Matrix - 3S Approach Validation Sample, $\tilde{A} = 1$

```
I_V1_Cont<-function(m,y,param) {

m3.11<-m[1]
m3.10<-m[2]
m3.01<-m[3]
m3.00<-m[4]

y3.11<-y[1]
y3.10<-y[2]
y3.01<-y[3]
y3.00<-y[4]

P11_mle<-param[1]
P10_mle<-param[2]

Phi11_mle<-param[3]
Phi10_mle<-param[4]
Phi01_mle<-param[5]
Phi00_mle<-param[6]
Vone11 <- 0;
Vone22 <-0;
Vone33 <- y3.11/Phi11_mle^2 - m3.11/(Phi11_mle + Phi01_mle)^2;
Vone44 <- y3.01/Phi10_mle^2 -m3.01/(Phi10_mle + Phi00_mle)^2;
Vone55 <- (m3.11 - y3.11)/Phi01_mle^2 - m3.11/(Phi11_mle + Phi01_mle)^2;
Vone66 <- (m3.01 - y3.01)/Phi00_mle^2 - m3.01/(Phi10_mle + Phi00_mle)^2;

Vone12 <- 0;
Vone13 <- 0;
Vone14 <- 0;
Vone15 <- 0;
Vone16 <- 0;
Vone23 <- 0;
Vone24 <- 0;
Vone25 <- 0;
Vone26 <- 0;
Vone34 <- 0;
Vone35 <- -m3.11/(Phi11_mle + Phi01_mle)^2;
Vone36 <- 0;
Vone45 <- 0;
Vone46 <- -m3.01/(Phi10_mle + Phi00_mle)^2;
Vone56 <- 0;

IV1<-rbind(c(Vone11, Vone12, Vone13, Vone14, Vone15, Vone16),
c(Vone12, Vone22, Vone23, Vone24, Vone25, Vone26),
c(Vone13, Vone23, Vone33, Vone34, Vone35, Vone36),
c(Vone14, Vone24, Vone34, Vone44, Vone45, Vone46),
c(Vone15, Vone25, Vone35, Vone45, Vone55, Vone56),
c(Vone16, Vone26, Vone36, Vone46, Vone56, Vone66))
return(IV1)
}
```

Information Matrix - 3S Approach Validation Sample, $\tilde{A} = 0$

```
I_V2_Cont<-function(m,y,param) {

m3.11<-m[1]
m3.10<-m[2]
m3.01<-m[3]
m3.00<-m[4]

y3.11<-y[1]
y3.10<-y[2]
y3.01<-y[3]
y3.00<-y[4]

P11_mle<-param[1]
P10_mle<-param[2]

Phi11_mle<-param[3]
Phi10_mle<-param[4]
Phi01_mle<-param[5]
Phi00_mle<-param[6]

Vtwo11 <- (m3.10 - y3.10)/(1 - P11_mle - Phi01_mle)^2 + y3.10/(P11_mle - Phi11_mle)^2;
Vtwo22 <- (m3.00 - y3.00)/(1 - P10_mle - Phi00_mle)^2 + y3.00/(P10_mle - Phi10_mle)^2;
Vtwo33 <- y3.10/(P11_mle - Phi11_mle)^2 - m3.10/(1 - Phi11_mle - Phi01_mle)^2 ;
Vtwo44 <- y3.00/(P10_mle - Phi10_mle)^2 - m3.00/(1 - Phi10_mle - Phi00_mle)^2;
Vtwo55 <- (m3.10 - y3.10)/(1 - P11_mle - Phi01_mle)^2 - m3.10/(1 - Phi11_mle - Phi01_mle)^2;
Vtwo66 <- (m3.00 - y3.00)/(1 - P10_mle - Phi00_mle)^2 - m3.00/(1 - Phi10_mle - Phi00_mle)^2 ;

Vtwo12 <- 0;
Vtwo13 <- -(y3.10/(P11_mle - Phi11_mle)^2);
Vtwo14 <- 0;
Vtwo15 <- (m3.10 - y3.10)/(1 - P11_mle - Phi01_mle)^2;
Vtwo16 <- 0;
Vtwo23 <- 0;
Vtwo24 <- -(y3.00/(P10_mle - Phi10_mle)^2);
Vtwo25 <- 0;
Vtwo26 <- (m3.00 - y3.00)/(1 - P10_mle - Phi00_mle)^2;
Vtwo34 <- 0;
Vtwo35 <- -m3.10/(1 - Phi11_mle - Phi01_mle)^2;
Vtwo36 <- 0;
Vtwo45 <- 0;
Vtwo46 <- -m3.00/(1 - Phi10_mle - Phi00_mle)^2;
Vtwo56 <- 0;

IV2<-rbind(c(Vtwo11, Vtwo12, Vtwo13, Vtwo14, Vtwo15, Vtwo16),
c(Vtwo12, Vtwo22, Vtwo23, Vtwo24, Vtwo25, Vtwo26),
c(Vtwo13, Vtwo23, Vtwo33, Vtwo34, Vtwo35, Vtwo36),
c(Vtwo14, Vtwo24, Vtwo34, Vtwo44, Vtwo45, Vtwo46),
c(Vtwo15, Vtwo25, Vtwo35, Vtwo45, Vtwo55, Vtwo56),
c(Vtwo16, Vtwo26, Vtwo36, Vtwo46, Vtwo56, Vtwo66))

return(IV2)
}
```

MLEs

```
#   2S approach

Phi11.2<-function(y11,m11,n11,n10) (y11*(n11+m11))/(m11*(n11 + n10 + m11 + m10))

Phi10.2<-function(y01,m01,n01,n00) (y01*(n01+m01))/(m01*(n01 + n00 + m01 + m00))

Phi01.2<-function(y11,m11,n11,n10) ((m11-y11)*(n11+m11))/(m11*(n11 + n10 + m11 + m10))
```

```
Phi00.2<-function(y01,m01,n01,n00) ((m01-y01)*(n01+m01))/(m01*(n01 + n00 + m01 + m00))

P11.2<-function(y11,y10,m11,m10,n11,n10)
(m11*n10*y10 + m10*(n11*y11+m11*(y10+y11)))/(m10*m11*(n11 + n10+m10+m11))

P10.2<-function(y01,y00,m01,m00,n01,n00)
(m01*n00*y00 + m00*(n01*y01+m01*(y00+y01)))/(m00*m01*(n01 + n00 + m00 + m01))


#   3S approach

Phi11.3<-function(y11,m11,n11,n10) (y11*n11)/(m11*(n11 + n10))

Phi10.3<-function(y01,m01,n01,n00) (y01*n01)/(m01*(n01 + n00))

Phi01.3<-function(y11,m11,n11,n10) (n11*(m11 - y11))/(m11*(n11 + n10))

Phi00.3<-function(y01,m01,n01,n00) (n01*(m01 - y01))/(m01*(n01 + n00))

P11.3<-function(y11,y10,m11,m10,n11,n10) (m11*n10*y10 + m10*n11*y11)/(m10*m11*(n11 + n10))

P10.3<-function(y01,y00,m01,m00,n01,n00) (m01*n00*y00 + m00*n01*y01)/(m00*m01*(n01 + n00))
```

IVW Estimator

```
IVW<-function(a,b1,b0)     {
#a, b1, b0 are of the form:
#a=c(a11,a01,a10,a00)
#b1=c(b111,b011,b101,b001)
#b0=c(b110,b010,b100,b000)
#
# where b_{a,\tilde{a},d} is the validation data
# and a_{\tilde{a},d} is from \{ S_V^C \cap S \}

a11<-a[1]
a01<-a[2]
a10<-a[3]
a00<-a[4]

b111<-b1[1]
b011<-b1[2]
b101<-b1[3]
b001<-b1[4]

b110<-b0[1]
b010<-b0[2]
b100<-b0[3]
b000<-b0[4]


Q1<-rbind(c(b111/(b111+b101),b011/(b011+b001)),c(b101/(b101+b111),b001/(b011+b001)))
Q0<-rbind(c(b110/(b110+b100),b010/(b000+b010)),c(b100/(b110+b100),b000/(b010+b000)))

Q<-rbind(cbind(Q1,matrix(0,2,2)),cbind(matrix(0,2,2),Q0))

c<-solve(Q)%*%a

vQ<-
(a11*a01*(a11+a01))/(c[1]*c[2]*det(Q1))^2+(a10*a00*(a10+a00))/(c[3]*c[4]*det(Q0))^2

vU<-((a11+a01)^2*b111/(b111+b101)^2*(1-b111/(b111+b101)))/(c[2]*det(Q1))^2 +
((a10+a00)^2*b110/(b110+b100)^2*(1-b110/(b110+b100)))/(c[4]*det(Q0))^2 +
((a11+a01)^2*b011/(b011+b001)^2*(1-b011/(b011+b001)))/(c[1]*det(Q1))^2+
((a10+a00)^2*b010/(b010+b000)^2*(1-b010/(b010+b000)))/(c[3]*det(Q0))^2

gamma_Qa<-log(c[1]*c[4]/(c[2]*c[3]))

V_Qa<-vQ+vU

gamma_b<-log((b111+b101)*(b010+b000)/((b110+b100)*(b011+b001)))
V_b<-1/(b111+b101)+1/(b010+b000)+1/(b110+b100)+1/(b011+b001)
V_Wa<-(1/V_Qa+1/V_b)^(-1)

gamma_Wa<-V_Wa*(gamma_Qa/V_Qa+gamma_b/V_b)

list(gamma_Wa=gamma_Wa,V_Wa=V_Wa,other=c(gamma_Qa,gamma_b))
}
```

Main Simulation Code

```
#################
#
# Code to generate Section 2 Tables
#

rm(list=ls(all=TRUE))

source("E:\\Thesis\\Ch. 1 OR\\Code\\FINAL\\Contingency Table Functions FINAL 12-10-2014.r")

numsim<-1000

N1<-5000
N0<-5000

N<-N1+N0

P11<-.2
P10<-.3333

ThetaMat<-c(.99,.95,.01,.05)

M<-1000

Natil<-(P11*ThetaMat[1]+(1-P11)*ThetaMat[3])*N1+(P10*ThetaMat[2]+(1-P10)*ThetaMat[4])*N0

OR<-P11*(1-P10)/(P10*(1-P11))

P1<-seq(.1,.9,.01)

ORtilda<-{}
LORtilda<-{}
VORtilda<-{}

IVW_OR<-{}
IVW_Var<-{}

res2mle<-{}
res3mle<-{}
Vars2<-{}
Vars3<-{}

LOR.2S<-{}
OR.2S<-{}
VOR.2S<-{}

LOR.3S<-{}
OR.3S<-{}
VOR.3S<-{}
NN1<-{}

SeedS<-{}
for (B in 1:numsim) {


DataSeed<-round(runif(1,1,1000000))
set.seed(DataSeed)
data<-datagenCont3(N1,N0,P11,P10,ThetaMat[1],ThetaMat[2],ThetaMat[3],ThetaMat[4])

Val2Seed<-round(runif(1,1,1000000))
set.seed(Val2Seed)
v2S<-Val2Sample(M,data$D)
```

```
#MLEs 2 sample
NN1<-c(NN1,sum(data$D[,3]))

n<-data$n
y.2<-v2S$y
m.2<-v2S$m

n11<-n[1]
n10<-n[2]
n01<-n[3]
n00<-n[4]

ORtilda<-c(ORtilda, n11*n00/(n10*n01))
LORtilda<-c(LORtilda, log(n11*n00/(n10*n01)))
VORtilda<-c(VORtilda,1/n11+1/n10+1/n01+1/n00)

y11<-y.2[1]
y10<-y.2[2]
y01<-y.2[3]
y00<-y.2[4]

m11<-m.2[1]
m10<-m.2[2]
m01<-m.2[3]
m00<-m.2[4]

# res2mle is of the form (P11,P10,Phi11,Phi10,Phi01,Phi00)
res2mle<-rbind(res2mle,c(P11.2(y11,y10,m11,m10,n11,n10),
P10.2(y01,y00,m01,m00,n01,n00),Phi11.2(y11,m11,n11,n10),
Phi10.2(y01,m01,n01,n00),Phi01.2(y11,m11,n11,n10),Phi00.2(y01,m01,n01,n00)))
CovMatrix<-solve(I_O_Cont3(n,res2mle[B,])+I_V_Cont3(m.2,y.2,res2mle[B,]))
Vars2<-rbind(Vars2,diag(CovMatrix))

#Odds ratio results
OR.2S<-c(OR.2S,res2mle[B,1]*(1-res2mle[B,2])/(res2mle[B,2]*(1-res2mle[B,1])))
LOR.2S<-c(LOR.2S,log(res2mle[B,1]*(1-res2mle[B,2])/(res2mle[B,2]*(1-res2mle[B,1]))))
VOR.2S<-c(VOR.2S,t(gOR(res2mle[B,1],res2mle[B,2],4))%*%CovMatrix%*%(gOR(res2mle[B,1],res2mle[B,2],4)))

#IVW results
IVWTemp<-IVW(c(n11-m11,n10-m10,n01-m01,n00-m00),c(y11,m11-y11,y10,m10-y10),c(y01,m01-y01,y00,m00-y00))
IVW_OR<-c(IVW_OR,IVWTemp$gamma_Wa)
IVW_Var<-c(IVW_Var,IVWTemp$V_Wa)


for (v in 1:length(P1)) {

M1<-round(P1[v]*M)

M0<-M-M1

Val3Seed<-round(runif(1,1,1000000))
set.seed(Val3Seed)
v3S<-Val3Sample(M1,M0,data$D)

y.3<-v3S$y
m.3<-v3S$m

y11<-y.3[1]
y10<-y.3[2]
y01<-y.3[3]
y00<-y.3[4]

m11<-m.3[1]
m10<-m.3[2]
m01<-m.3[3]
m00<-m.3[4]

res3mle<-rbind(res3mle,c(P11.3(y11,y10,m11,m10,n11,n10),
P10.3(y01,y00,m01,m00,n01,n00),Phi11.3(y11,m11,n11,n10),
Phi10.3(y01,m01,n01,n00),Phi01.3(y11,m11,n11,n10),Phi00.3(y01,m01,n01,n00)))

CovMatrix<-solve(I_O_Cont3(n,res3mle[((B-1)*length(P1)+v),])+
I_V1_Cont3(m.3,y.3,res3mle[((B-1)*length(P1)+v),])+I_V2_Cont3(m.3,y.3,res3mle[((B-1)*length(P1)+v),]))
Vars3<-rbind(Vars3,diag(CovMatrix))

#Odds ratio results
OR.3S<-c(OR.3S,res3mle[((B-1)*length(P1)+v),1]*(1-res3mle[((B-1)*length(P1)+v),2])/
(res3mle[((B-1)*length(P1)+v),2]*(1-res3mle[((B-1)*length(P1)+v),1])))
LOR.3S<-c(LOR.3S,log(res3mle[((B-1)*length(P1)+v),1]*(1-res3mle[((B-1)*length(P1)+v),2])/
(res3mle[((B-1)*length(P1)+v),2]*(1-res3mle[((B-1)*length(P1)+v),1]))))
VOR.3S<-c(VOR.3S,t(gOR(res3mle[((B-1)*length(P1)+v),1],res3mle[((B-1)*length(P1)+v),2],4))%*%CovMatrix%*%
(gOR(res3mle[((B-1)*length(P1)+v),1],res3mle[((B-1)*length(P1)+v),2],4)))

}
}

#The 3S results are ordered by row as P1[1],P1[2],... for each iteration
colnames(res2mle)<-c("P11","P10","Phi11","Phi10","Phi01","Phi00")
colnames(res3mle)<-c("P11","P10","Phi11","Phi10","Phi01","Phi00")
colnames(Vars2)<-c("P11","P10","Phi11","Phi10","Phi01","Phi00")
colnames(Vars3)<-c("P11","P10","Phi11","Phi10","Phi01","Phi00")


################
#
# MLEs and estimates of error
#
#

PE.3S<-{}
Var.3S<-{}
VarMC.3S<-{}
OR3results<-{}
VOR3results<-{}
OR3results.MC<-{}

OR3CIresults<-{}


for(k in 1:length(P1)) {
PE.3S<-rbind(PE.3S,apply(res3mle[seq(k,dim(res3mle)[1],length(P1)),],2,mean))
Var.3S<-rbind(Var.3S,apply(Vars3[seq(k,dim(Vars3)[1],length(P1)),],2,mean))
VarMC.3S<-rbind(VarMC.3S,apply(res3mle[seq(k,dim(res3mle)[1],length(P1)),],2,var))

OR3results<-c(OR3results,mean(OR.3S[seq(k,dim(res3mle)[1],length(P1))]))
VOR3results<-c(VOR3results,mean(VOR.3S[seq(k,dim(res3mle)[1],length(P1))]))
OR3results.MC<-c(OR3results.MC,var(log(OR.3S[seq(k,dim(res3mle)[1],length(P1))])))


OR3CIresults<-c(OR3CIresults,length(which(log(OR)>log(OR.3S[seq(k,dim(res3mle)[1],
length(P1))])-1.96*sqrt(VOR.3S[seq(k,dim(res3mle)[1],length(P1))])
&log(OR)<log(OR.3S[seq(k,dim(res3mle)[1],length(P1))])+
1.96*sqrt(VOR.3S[seq(k,dim(res3mle)[1],length(P1))])))/numsim)

}

rownames(PE.3S)<-as.character(P1)
rownames(Var.3S)<-as.character(P1)
rownames(VarMC.3S)<-as.character(P1)


PE.2S<-apply(res2mle,2,mean)
Var.2S<-apply(Vars2,2,mean)
VarMC.2S<-apply(res2mle,2,var)

################
#
# PE.2S and PE.3S are the point estimates, Var.2S and Var.3S are the variance estiamtes
# VarMC.2S and VarMC.3S are the simulation error estimates
# The 3S approach results are set up such that the rows are the different values of P1.
#
##################

################
#
# SE tables, 2.5 - 2.7
```

```
#

SEresults<-round(cbind(c(sqrt(Var.2S[1:2]),max(sqrt(Var.2S[1:2]))),
sqrt(rbind(t(Var.3S[,1:2]),apply(t(Var.3S[,1:2]),2,max)))),5)
rownames(SEresults)<-c("SE(P11)","SE(P10)","max")
colnames(SEresults)<-c("2 samp",as.character(P1))


#
#
################

################
#
# Odds Ratio Tables - assumes all P1 values are run as seq(0.1,0.9,0.1)
#

P1incidenceRate.index<-which(P1==round(mean(NN1)/N,2))



P1min.index<-which(SEresults[3,]==min(SEresults[3,]))


ORresults<-cbind(
c(mean(ORtilda),mean(OR.2S),OR3results[P1min.index],OR3results[P1incidenceRate.index],mean(exp(IVW_OR))),
c(mean(sqrt(VORtilda)),mean(sqrt(VOR.2S)),sqrt(VOR3results[P1min.index]),
sqrt(VOR3results[P1incidenceRate.index]),mean(sqrt(IVW_Var))),c(sqrt(var(log(ORtilda))),
sqrt(var(log(OR.2S))),sqrt(OR3results.MC[P1min.index]),
sqrt(OR3results.MC[P1incidenceRate.index]),sqrt(var(IVW_OR))),
c(length(which(log(OR)>LORtilda-1.96*sqrt(VORtilda)&log(OR)<LORtilda+1.96*sqrt(VORtilda)))/numsim,
length(which(log(OR)>log(OR.2S)-1.96*sqrt(VOR.2S)&log(OR)<log(OR.2S)+1.96*sqrt(VOR.2S)))/numsim,
OR3CIresults[P1min.index],OR3CIresults[P1incidenceRate.index],
length(which(log(OR)>IVW_OR-1.96*sqrt(IVW_Var)&log(OR)<IVW_OR+1.96*sqrt(IVW_Var)))/numsim))

rownames(ORresults)<-c("ORtil","ORhat2","ORhat3.min","ORhat3.in","ORivw")
colnames(ORresults)<-c("PE","SE","SEMC","CI")
#
#
################


################
#
# Interval Results - see Ch. 2 Tables 2.11 and on
#

LB2S<-ifelse(length(P1[which(diff(sign(mean(VOR.2S)-VOR3results))>0)])!=0,
P1[which(diff(sign(mean(VOR.2S)-VOR3results))>0)],-1)
UB2S<-ifelse(length(P1[which(diff(sign(mean(VOR.2S)-VOR3results))<0)])!=0,
P1[which(diff(sign(mean(VOR.2S)-VOR3results))<0)],-1)

LBivw<-ifelse(length(P1[which(diff(sign(mean(IVW_Var)-VOR3results))>0)])!=0,
P1[which(diff(sign(mean(IVW_Var)-VOR3results))>0)],-1)
UBivw<-ifelse(length(P1[which(diff(sign(mean(IVW_Var)-VOR3results))<0)])!=0,
P1[which(diff(sign(mean(IVW_Var)-VOR3results))<0)],-1)
INTresults<-c(Natil/N,OR,P11,P10,ThetaMat,LB2S,UB2S,LBivw,UBivw,mean(VOR.2S),mean(IVW_Var),min(VOR3results))
```

C.2 Chapter 3 Code

Sample Size Determination

```
InfoApprox<-function(N1,N0,parList,numIter=1000,P_m=0.3,P1gen=1) {
#parList is of the form (pi11,pi10,th11,th01,th10,th00)

N<-N1+N0
M<-P_m*N

#######Initialize##########

res<-{}
NN1<-{}
SeedS<-{}

IVMC<-matrix(0,6,6)
IOMC<-matrix(0,6,6)
IV1MC<-matrix(0,6,6)
IV2MC<-matrix(0,6,6)

##########################

P11<-parList[1]
P10<-parList[2]
Th11<-parList[3]
Th01<-parList[4]
Th10<-parList[5]
Th00<-parList[6]

Natil<-(P11*Th11+(1-P11)*Th10)*N1+(P10*Th01+(1-P10)*Th00)*N0
M1<-round((Natil/N)*P1gen*M)
M2<-M-M1

nn1<-{}
for (i in 1:numIter) {

tempseedi<-runif(1,1,10000000)
set.seed(tempseedi)
data<-datagenCont3(N1,N0,P11,P10,Th11,Th01, Th10, Th00)

n11<-data$n[1]
n10<-data$n[2]
n01<-data$n[3]
n00<-data$n[4]

nn1<-c(nn1,sum(data$D[,3]))

Phi11<-Th11*P11
Phi10<-Th01*P10
Phi01<-Th10*(1-P11)
Phi00<-Th00*(1-P10)

IO<-I_O_Cont3(data$n,c(P11,P10,Phi11,Phi10,Phi01,Phi00))

tempseedV2<-runif(1,1,10000000)
set.seed(tempseedV2)
v2S<-Val2Sample(M,data$D)

tempseedV3<-runif(1,1,10000000)
set.seed(tempseedV3)
v3S<-Val3Sample(M1,M2,data$D)

IV<-I_V_Cont3(v2S$m,v2S$y,c(P11,P10,Phi11,Phi10,Phi01,Phi00))
IV1<-I_V1_Cont3(v3S$m,v3S$y,c(P11,P10,Phi11,Phi10,Phi01,Phi00))
IV2<-I_V2_Cont3(v3S$m,v3S$y,c(P11,P10,Phi11,Phi10,Phi01,Phi00))

IOMC<-IOMC+(IO/N)
IVMC<-IVMC+(IV/M)
IV1MC<-IV1MC+(IV1/M1)
IV2MC<-IV2MC+(IV2/M2)
SeedS<-rbind(SeedS,c(tempseedi,tempseedV2,tempseedV3))
colnames(SeedS)<-c("O","V2samp","V3samp")
}

list(Io=IOMC/numIter,Iv=IVMC/numIter,Iv1=IV1MC/numIter,Iv2=IV2MC/numIter,n1=mean(nn1),SeedS=SeedS)
}


SSdet<-function(N1,N0,parList,BoundSeq=seq(0.01,0.4,0.01),P_m=0.3,P1gen=1,numIter=1000,Seed=FALSE) {

##Init
res11<-{}
res10<-{}
SampleSizeTable<-{}
###

N<-N1+N0

INFO<-InfoApprox(N1,N0,parList,numIter,P_m,P1gen)
n1<-INFO$n1
Io<-INFO$Io
Iv<-INFO$Iv
Iv1<-INFO$Iv1
Iv2<-INFO$Iv2
p<-dim(Io)[1]

b<-rbind(c(1,rep(0,p-1)),c(0,1,rep(0,p-2)))

f2<-function(m,B,b) qnorm(1-.05/(2*p))*sqrt(t(b)%*%solve(N*Io+m*Iv)%*%b)-B

for(i in 1:length(BoundSeq)){
res11<-c(res11,tryCatch(ceiling(uniroot(f2,c(1,N),B=BoundSeq[i],b=b[1,])$root),error=function(e) -1))
res10<-c(res10,tryCatch(ceiling(uniroot(f2,c(1,N),B=BoundSeq[i],b=b[2,])$root),error=function(e) -1))
}
res2samp<-cbind(pmax(res11,res10),BoundSeq)

P1.3<-seq(.1,.9,(.9-.1)/8)
f3<-function(m,B,pv,b) qnorm(1-.05/(2*p))*sqrt(t(b)%*%solve(N*Io+m*(pv*Iv1+(1-pv)*Iv2))%*%b)-B

res3<-{}

for(i in 1:length(BoundSeq))
for(j in 1:length(P1.3)) {
#If one of the uniroot calls produces an error due to sign change,
#that implies that the max would be larger than N hence, no value is returned
evalMax<-tryCatch(max(ceiling(uniroot(f3,c(1,N),B=BoundSeq[i],pv=P1.3[j],b[1,])$root),
ceiling(uniroot(f3,c(5,N),B=BoundSeq[i],pv=P1.3[j],b[2,])$root)),error=function(e) -1)
res3<-rbind(res3,c(evalMax,P1.3[j],BoundSeq[i]))
}

res3[which(res3[,1]*res3[,2]>n1),1]<--1

temp3<-SortByP1(res3)
SSdet<-cbind(BoundSeq,matrix(temp3[,1],nrow=length(BoundSeq)),res2samp[,1])
colnames(SSdet)<-c("c","0.1","0.2","0.3","0.4","0.5","0.6","0.7","0.8","0.9","2 samp")
SampleSizeTable$SSdet<-SSdet
if(Seed==TRUE) SampleSizeTable$SeedS=INFO$SeedS
return(SampleSizeTable)
}
```

C.3 Chapter 4 Code

Data Generation

```
datagenBinary<-function(n1,n0,alpha,beta,gamma1,gamma2,kappa11,
kappa01,kappa10,kappa00,tau1,tau2,Pz,parZ2=c(2,1)) {

z1_1<-rbinom(n1,1,Pz)
z1_0<-rbinom(n0,1,Pz)

z2_1<-rgamma(n1,parZ2)
z2_0<-rgamma(n0,parZ2)

#probabilities of form Pei, e=0,1, zi=0,1
P1<-unpack(alpha+beta+gamma1*z1_1+gamma2*z2_1)
P0<-unpack(alpha+gamma1*z1_0+gamma2*z2_0)

Y1<-rbinom(n1,1,P1)
Y0<-rbinom(n0,1,P0)

Atil1<-rep(0,n1)
Atil0<-rep(0,n0)

#The1i ie only a tilda = 1

Th11<-Y1*unpack(kappa11+tau1*z1_1+tau2*z2_1)
Th10<-(1-Y1)*unpack(kappa10+tau1*z1_1+tau2*z2_1)
Th01<-Y0*unpack(kappa01+tau1*z1_0+tau2*z2_0)
Th00<-(1-Y0)*unpack(kappa00+tau1*z1_0+tau2*z2_0)

Obs1_1<-rbinom(sum(Y1),1,Th11[which(Th11!=0)])
Obs1_0<-rbinom(n1-sum(Y1),1,Th10[which(Th10!=0)])

Atil1[which(Y1==1)]<-Obs1_1
Atil1[which(Y1==0)]<-Obs1_0

#observed as a 0
Obs0_1<-rbinom(sum(Y0),1,Th01[which(Th01!=0)])
Obs0_0<-rbinom(n0-sum(Y0),1,Th00[which(Th00!=0)])

Atil0[which(Y0==1)]<-Obs0_1
Atil0[which(Y0==0)]<-Obs0_0

D<-c(rep(1,n1),rep(0,n0))
Z1<-c(z1_1,z1_0)
Z2<-c(z2_1,z2_0)
A<-c(Y1,Y0)
Atilda<-c(Atil1,Atil0)

dataset<-cbind(D,A,Z1,Z2,Atilda)

return(dataset)
}
```

2S Approach: Validation Sampling Algorithm

```
Val2SampleBinary<-function(D,M) {      #2 sample approach validation sample algorithm

ValInd<-sample(seq(1,length(D[,5])),M)
ValComp<-setdiff(seq(1,length(D[,5])),ValInd)

Validation<-D[ValInd,]
ValComp<-D[ValComp,]
list(Validation=Validation,ValComp=ValComp)

}
```

3S Approach: Validation Sampling Algorithm

```
Val3SampleBinary<-function(D,M1,M2) { #3 sample approach validation sample algorithm
#Check to see if enough were sampled in each group?
Atilde1<-which(D[,5]==1)
Atilde2<-which(D[,5]==0)

ValInd1<-sample(seq(1,length(Atilde1)),M1)
ValInd2<-sample(seq(1,length(Atilde2)),M2)
ValComp1<-setdiff(seq(1,length(Atilde1)),ValInd1)
ValComp2<-setdiff(seq(1,length(Atilde2)),ValInd2)

Validation<-rbind(D[Atilde1,][ValInd1,],D[Atilde2,][ValInd2,])
ValComp<-rbind(D[Atilde1,][ValComp1,],D[Atilde2,][ValComp2,])

list(Validation=Validation,ValComp=ValComp)

}
```

Information Matrix - Original Sample

```
I0<-function(data,param) {

D<-data[,1]
A<-data[,2]
Z1<-data[,3]
Z2<-data[,4]
ATIL<-data[,5]

a<-param[1]
b<-param[2]
g1<-param[3]
g2<-param[4]
k11<-param[5]
k01<-param[6]
k10<-param[7]
k00<-param[8]
t1<-param[9]
t2<-param[10]

eta<-exp(a+b*D+g1*Z1+g2*Z2)
rho1<-exp(k11*D+k01*(1-D)+t1*Z1+t2*Z2)
rho0<-exp(k10*D+k00*(1-D)+t1*Z1+t2*Z2)

dETA.ETA<-eta/(1+eta)^2-((1-ATIL)*eta*(1+rho1)*(1+rho0))/(1+eta+rho1+eta*rho0)^2-
(ATIL*eta*rho1*rho0*(1+rho1)*(1+rho0))/(rho0+rho1*rho0+eta*rho1+eta*rho1*rho0)^2

dRHO1.RHO1<-rho1/(1+rho1)^2- ((1-ATIL)*rho1*(1+eta+eta*rho0))/(1+eta+rho1+eta*rho0)^2-
(ATIL*rho1*rho0*(eta+rho0+eta*rho0))/(rho0+rho1*rho0+eta*rho1+eta*rho1*rho0)^2

dRHO0.RHO0<-rho0/(1+rho0)^2- ((1-ATIL)*eta*rho0*(1+eta+rho1))/(1+eta+rho1+eta*rho0)^2-
(ATIL*eta*rho1*rho0*(1+rho1+eta*rho1))/(rho0+rho1*rho0+eta*rho1+eta*rho1*rho0)^2

dETA.RHO1<- ((1-ATIL)*rho1*eta*(1+rho0))/(1+eta+rho1+eta*rho0)^2-
(ATIL*eta*rho1*rho0*(1+rho0))/(rho0+rho1*rho0+eta*rho1+eta*rho1*rho0)^2

dETA.RHO0<- (ATIL*eta*rho1*rho0*(1+rho1))/(rho0+rho1*rho0+eta*rho1+eta*rho1*rho0)^2 -
((1-ATIL)*rho0*eta*(1+rho1))/(1+eta+rho1+eta*rho0)^2

dRHO1.RHO0<- (ATIL*eta*rho1*rho0)/(rho0+rho1*rho0+eta*rho1+eta*rho1*rho0)^2 +
((1-ATIL)*eta*rho1*rho0)/(1+eta+rho1+eta*rho0)^2

I11<-sum(dETA.ETA)

I22<-sum(D^2*dETA.ETA)

I33<-sum(Z1^2*dETA.ETA)

I44<- sum(Z2^2*dETA.ETA)

I55<- sum(D^2*dRHO1.RHO1)

I66<- sum((1-D)^2*dRHO1.RHO1)
```

```
I77<-sum(D^2*dRHO0.RHO0)

I88<-sum((1-D)^2*dRHO0.RHO0)

I99<- sum(Z1^2*(dRHO1.RHO1+dRHO0.RHO0+2*dRHO1.RHO0))

I1010<-sum(Z2^2*(dRHO1.RHO1+dRHO0.RHO0+2*dRHO1.RHO0))

I12<-sum(D*dETA.ETA)

I13<-sum(Z1*dETA.ETA)

I14<-sum(Z2*dETA.ETA)

I15<-sum(D*dETA.RHO1)

I16<-sum((1-D)*dETA.RHO1)

I17<-sum(D*dETA.RHO0)

I18<-sum((1-D)*dETA.RHO0)

I19<- sum(Z1*(dETA.RHO1+dETA.RHO0))

I110<-sum(Z2*(dETA.RHO1+dETA.RHO0))

I23<-sum(D*Z1*dETA.ETA)

I24<-sum(D*Z2*dETA.ETA)

I25<-sum(D^2*dETA.RHO1)

I26<-0

I27<-sum(D^2*dETA.RHO0)

I28<-0

I29<-sum(D*Z1*(dETA.RHO1+dETA.RHO0))

I210<-sum(D*Z2*(dETA.RHO1+dETA.RHO0))

I34<-sum(Z1*Z2*dETA.ETA)

I35<-sum(Z1*D*dETA.RHO1)

I36<-sum(Z1*(1-D)*dETA.RHO1)

I37<-sum(Z1*D*dETA.RHO0)

I38<-sum(Z1*(1-D)*dETA.RHO0)

I39<-sum(Z1^2*(dETA.RHO1+dETA.RHO0))

I310<-sum(Z1*Z2*(dETA.RHO1+dETA.RHO0))

I45<-sum(Z2*D*dETA.RHO1)

I46<-sum(Z2*(1-D)*dETA.RHO1)

I47<-sum(Z2*D*dETA.RHO0)

I48<-sum(Z2*(1-D)*dETA.RHO0)

I49<-sum(Z1*Z2*(dETA.RHO1+dETA.RHO0))

I410<-sum(Z2^2*(dETA.RHO1+dETA.RHO0))

I56<-0

I57<-sum(D^2*dRHO1.RHO0)

I58<-0
```

```
I59<-sum(Z1*D*(dRHO1.RHO1+dRHO1.RHO0))

I510<-sum(Z2*D*(dRHO1.RHO1+dRHO1.RHO0))

I67<-0

I68<-sum((1-D)^2*dRHO1.RHO0)

I69<-sum(Z1*(1-D)*(dRHO1.RHO1+dRHO1.RHO0))

I610<-sum(Z2*(1-D)*(dRHO1.RHO1+dRHO1.RHO0))

I78<-0

I79<-sum(Z1*D*(dRHO0.RHO0+dRHO1.RHO0))


I710<-sum(Z2*D*(dRHO0.RHO0+dRHO1.RHO0))


I89<-sum(Z1*(1-D)*(dRHO0.RHO0+dRHO1.RHO0))

I810<-sum(Z2*(1-D)*(dRHO0.RHO0+dRHO1.RHO0))

I910<-sum(Z1*Z2*(dRHO1.RHO1+dRHO0.RHO0+2*dRHO1.RHO0))


I<-rbind(c(I11,I12,I13,I14,I15,I16,I17,I18,I19,I110),
         c(I12,I22,I23,I24,I25,I26,I27,I28,I29,I210),
         c(I13,I23,I33,I34,I35,I36,I37,I38,I39,I310),
c(I14,I24,I34,I44,I45,I46,I47,I48,I49,I410),
c(I15,I25,I35,I45,I55,I56,I57,I58,I59,I510),
c(I16,I26,I36,I46,I56,I66,I67,I68,I69,I610),
c(I17,I27,I37,I47,I57,I67,I77,I78,I79,I710),
c(I18,I28,I38,I48,I58,I68,I78,I88,I89,I810),
c(I19,I29,I39,I49,I59,I69,I79,I89,I99,I910),
c(I110,I210,I310,I410,I510,I610,I710,I810,I910,I1010))

return(I)
}
```

Information Matrix - 2S Approach Validation Sample

```
IV<-function(data,param) {

D<-data[,1]
A<-data[,2]
Z1<-data[,3]
Z2<-data[,4]
ATIL<-data[,5]

a<-param[1]
b<-param[2]
g1<-param[3]
g2<-param[4]
k11<-param[5]
k01<-param[6]
k10<-param[7]
k00<-param[8]
t1<-param[9]
t2<-param[10]

eta<-exp(a+b*D+g1*Z1+g2*Z2)
rho1<-exp(k11*D+k01*(1-D)+t1*Z1+t2*Z2)
rho0<-exp(k10*D+k00*(1-D)+t1*Z1+t2*Z2)

dETA.ETA<- eta/(1+eta)^2

dRHO1.RHO1<-A*rho1/(1+rho1)^2
```

```
dRHO0.RHO0<-(1-A)*rho0/(1+rho0)^2

dETA.RHO1<-0

dETA.RHO0<-0

dRHO1.RHO0<-0

I11<-sum(dETA.ETA)

I22<-sum(D^2*dETA.ETA)

I33<-sum(Z1^2*dETA.ETA)

I44<-sum(Z2^2*dETA.ETA)

I55<-sum(D^2*dRHO1.RHO1)

I66<-sum((1-D)^2*dRHO1.RHO1)

I77<-sum(D^2*dRHO0.RHO0)

I88<-sum((1-D)^2*dRHO0.RHO0)

I99<-sum(Z1^2*(dRHO1.RHO1+dRHO0.RHO0+2*dRHO1.RHO0))

I1010<-sum(Z2^2*(dRHO1.RHO1+dRHO0.RHO0+2*dRHO1.RHO0))

I12<-sum(D*dETA.ETA)

I13<-sum(Z1*dETA.ETA)

I14<-sum(Z2*dETA.ETA)

I15<-sum(D*dETA.RHO1)

I16<-sum((1-D)*dETA.RHO1)

I17<-sum(D*dETA.RHO0)

I18<-sum((1-D)*dETA.RHO0)

I19<-sum(Z1*(dETA.RHO1+dETA.RHO0))

I110<-sum(Z2*(dETA.RHO1+dETA.RHO0))

I23<-sum(D*Z1*dETA.ETA)

I24<-sum(D*Z2*dETA.ETA)

I25<-sum(D^2*dETA.RHO1)

I26<-0

I27<-sum(D^2*dETA.RHO0)

I28<-0

I29<-sum(D*Z1*(dETA.RHO1+dETA.RHO0))

I210<-sum(D*Z2*(dETA.RHO1+dETA.RHO0))

I34<-sum(Z1*Z2*dETA.ETA)

I35<-sum(Z1*D*dETA.RHO1)

I36<-sum(Z1*(1-D)*dETA.RHO1)

I37<-sum(Z1*D*dETA.RHO0)

I38<-sum(Z1*(1-D)*dETA.RHO0)

I39<-sum(Z1^2*(dETA.RHO1+dETA.RHO0))
```

```
I310<-sum(Z1*Z2*(dETA.RHO1+dETA.RHO0))

I45<-sum(Z2*D*dETA.RHO1)

I46<-sum(Z2*(1-D)*dETA.RHO1)

I47<-sum(Z2*D*dETA.RHO0)

I48<-sum(Z2*(1-D)*dETA.RHO0)

I49<-sum(Z1*Z2*(dETA.RHO1+dETA.RHO0))

I410<-sum(Z2^2*(dETA.RHO1+dETA.RHO0))

I56<-0

I57<-sum(D^2*dRHO1.RHO0)

I58<-0

I59<-sum(Z1*D*(dRHO1.RHO1+dRHO1.RHO0))

I510<-sum(Z2*D*(dRHO1.RHO1+dRHO1.RHO0))

I67<-0

I68<-sum((1-D)^2*dRHO1.RHO0)


I69<-sum(Z1*(1-D)*(dRHO1.RHO1+dRHO1.RHO0))

I610<-sum(Z2*(1-D)*(dRHO1.RHO1+dRHO1.RHO0))

I78<-0

I79<-sum(Z1*D*(dRHO0.RHO0+dRHO1.RHO0))

I710<-sum(Z2*D*(dRHO0.RHO0+dRHO1.RHO0))

I89<-sum(Z1*(1-D)*(dRHO0.RHO0+dRHO1.RHO0))

I810<-sum(Z2*(1-D)*(dRHO0.RHO0+dRHO1.RHO0))

I910<-sum(Z1*Z2*(dRHO1.RHO1+dRHO0.RHO0+2*dRHO1.RHO0))

I<-rbind(c(I11,I12,I13,I14,I15,I16,I17,I18,I19,I110),
         c(I12,I22,I23,I24,I25,I26,I27,I28,I29,I210),
         c(I13,I23,I33,I34,I35,I36,I37,I38,I39,I310),
c(I14,I24,I34,I44,I45,I46,I47,I48,I49,I410),
c(I15,I25,I35,I45,I55,I56,I57,I58,I59,I510),
c(I16,I26,I36,I46,I56,I66,I67,I68,I69,I610),
c(I17,I27,I37,I47,I57,I67,I77,I78,I79,I710),
c(I18,I28,I38,I48,I58,I68,I78,I88,I89,I810),
c(I19,I29,I39,I49,I59,I69,I79,I89,I99,I910),
c(I110,I210,I310,I410,I510,I610,I710,I810,I910,I1010))

return(I)
}
```

Information Matrix - 3S Approach Validation Sample, $\tilde{A} = 1$

```
IV.1<-function(data,param) {

D<-data[,1]
A<-data[,2]
Z1<-data[,3]
Z2<-data[,4]
ATIL<-data[,5]

a<-param[1]
```

```
b<-param[2]
g1<-param[3]
g2<-param[4]
k11<-param[5]
k01<-param[6]
k10<-param[7]
k00<-param[8]
t1<-param[9]
t2<-param[10]

eta<-exp(a+b*D+g1*Z1+g2*Z2)
rho1<-exp(k11*D+k01*(1-D)+t1*Z1+t2*Z2)
rho0<-exp(k10*D+k00*(1-D)+t1*Z1+t2*Z2)

dETA.ETA<- (ATIL*eta*rho1*rho0*(1+rho1)*(1+rho0))/(rho0+rho1*rho0+eta*rho1+eta*rho1*rho0)^2

dRHO1.RHO1<-(ATIL*rho1*rho0*(eta+rho0+eta*rho0))/(rho0+rho1*rho0+eta*rho1+eta*rho1*rho0)^2-
(ATIL*(1-A)*rho1)/(1+rho1)^2

dRHO0.RHO0<-(ATIL*eta*rho1*rho0*(1+rho1+eta*rho1))/(rho0+rho1*rho0+eta*rho1+eta*rho1*rho0)^2-
(ATIL*A*rho0)/(1+rho0)^2

dETA.RHO1<-(ATIL*eta*rho1*rho0*(1+rho0))/(rho0+rho1*rho0+eta*rho1+eta*rho1*rho0)^2

dETA.RHO0<--(ATIL*eta*rho1*rho0*(1+rho1))/(rho0+rho1*rho0+eta*rho1+eta*rho1*rho0)^2

dRHO1.RHO0<--(ATIL*eta*rho1*rho0)/(rho0+rho1*rho0+eta*rho1+eta*rho1*rho0)^2

I11<-sum(dETA.ETA)

I22<-sum(D^2*dETA.ETA)

I33<-sum(Z1^2*dETA.ETA)

I44<- sum(Z2^2*dETA.ETA)

I55<- sum(D^2*dRHO1.RHO1)

I66<- sum((1-D)^2*dRHO1.RHO1)

I77<-sum(D^2*dRHO0.RHO0)

I88<-sum((1-D)^2*dRHO0.RHO0)

I99<-sum(Z1^2*(dRHO1.RHO1+dRHO0.RHO0+2*dRHO1.RHO0))

I1010<-sum(Z2^2*(dRHO1.RHO1+dRHO0.RHO0+2*dRHO1.RHO0))

I12<-sum(D*dETA.ETA)

I13<-sum(Z1*dETA.ETA)

I14<-sum(Z2*dETA.ETA)

I15<-sum(D*dETA.RHO1)

I16<-sum((1-D)*dETA.RHO1)

I17<-sum(D*dETA.RHO0)

I18<-sum((1-D)*dETA.RHO0)

I19<-sum(Z1*(dETA.RHO1+dETA.RHO0))

I110<-sum(Z2*(dETA.RHO1+dETA.RHO0))

I23<-sum(D*Z1*dETA.ETA)

I24<-sum(D*Z2*dETA.ETA)

I25<-sum(D^2*dETA.RHO1)

I26<-0
```

```
I27<-sum(D^2*dETA.RHO0)

I28<-0

I29<-sum(D*Z1*(dETA.RHO1+dETA.RHO0))

I210<-sum(D*Z2*(dETA.RHO1+dETA.RHO0))

I34<-sum(Z1*Z2*dETA.ETA)

I35<-sum(Z1*D*dETA.RHO1)

I36<-sum(Z1*(1-D)*dETA.RHO1)

I37<-sum(Z1*D*dETA.RHO0)

I38<-sum(Z1*(1-D)*dETA.RHO0)

I39<-sum(Z1^2*(dETA.RHO1+dETA.RHO0))

I310<-sum(Z1*Z2*(dETA.RHO1+dETA.RHO0))

I45<-sum(Z2*D*dETA.RHO1)

I46<-sum(Z2*(1-D)*dETA.RHO1)

I47<-sum(Z2*D*dETA.RHO0)

I48<-sum(Z2*(1-D)*dETA.RHO0)

I49<-sum(Z1*Z2*(dETA.RHO1+dETA.RHO0))

I410<-sum(Z2^2*(dETA.RHO1+dETA.RHO0))

I56<-0

I57<-sum(D^2*dRHO1.RHO0)

I58<-0

I59<-sum(Z1*D*(dRHO1.RHO1+dRHO1.RHO0))

I510<-sum(Z2*D*(dRHO1.RHO1+dRHO1.RHO0))

I67<-0

I68<-sum((1-D)^2*dRHO1.RHO0)

I69<-sum(Z1*(1-D)*(dRHO1.RHO1+dRHO1.RHO0))

I610<-sum(Z2*(1-D)*(dRHO1.RHO1+dRHO1.RHO0))

I78<-0

I79<-sum(Z1*D*(dRHO0.RHO0+dRHO1.RHO0))

I710<-sum(Z2*D*(dRHO0.RHO0+dRHO1.RHO0))

I89<-sum(Z1*(1-D)*(dRHO0.RHO0+dRHO1.RHO0))

I810<-sum(Z2*(1-D)*(dRHO0.RHO0+dRHO1.RHO0))

I910<-sum(Z1*Z2*(dRHO1.RHO1+dRHO0.RHO0+2*dRHO1.RHO0))

I<-rbind(c(I11,I12,I13,I14,I15,I16,I17,I18,I19,I110),
         c(I12,I22,I23,I24,I25,I26,I27,I28,I29,I210),
         c(I13,I23,I33,I34,I35,I36,I37,I38,I39,I310),
c(I14,I24,I34,I44,I45,I46,I47,I48,I49,I410),
c(I15,I25,I35,I45,I55,I56,I57,I58,I59,I510),
c(I16,I26,I36,I46,I56,I66,I67,I68,I69,I610),
c(I17,I27,I37,I47,I57,I67,I77,I78,I79,I710),
c(I18,I28,I38,I48,I58,I68,I78,I88,I89,I810),
c(I19,I29,I39,I49,I59,I69,I79,I89,I99,I910),
c(I110,I210,I310,I410,I510,I610,I710,I810,I910,I1010))

return(I)
}
```

Information Matrix - 3S Approach Validation Sample, $\tilde{A} = 0$

```
IV.2<-function(data,param) {

D<-data[,1]
A<-data[,2]
Z1<-data[,3]
Z2<-data[,4]
ATIL<-data[,5]

a<-param[1]
b<-param[2]
g1<-param[3]
g2<-param[4]
k11<-param[5]
k01<-param[6]
k10<-param[7]
k00<-param[8]
1<-param[9]
t2<-param[10]


eta<-exp(a+b*D+g1*Z1+g2*Z2)
rho1<-exp(k11*D+k01*(1-D)+t1*Z1+t2*Z2)
rho0<-exp(k10*D+k00*(1-D)+t1*Z1+t2*Z2)

dETA.ETA<-((1-ATIL)*eta*(1+rho1)*(1+rho0))/(1+eta+rho1+eta*rho0)^2

dRHO1.RHO1<-((1-ATIL)*rho1*(1+eta+eta*rho0))/(1+eta+rho1+eta*rho0)^2-
((1-A)*(1-ATIL)*rho1)/(1+rho1)^2

dRHO0.RHO0<-((1-ATIL)*eta*rho0*(1+eta+rho1))/(1+eta+rho1+eta*rho0)^2-
(A*(1-ATIL)*rho0)/(1+rho0)^2

dETA.RHO1<-(-(1-ATIL)*eta*rho1*(1+rho0))/(1+eta+rho1+eta*rho0)^2

dETA.RHO0<-((1-ATIL)*eta*rho0*(1+rho1))/(1+eta+rho1+eta*rho0)^2

dRHO1.RHO0<-(-(1-ATIL)*eta*rho0*rho1)/(1+eta+rho1+eta*rho0)^2

I11<-sum(dETA.ETA)

I22<-sum(D^2*dETA.ETA)

I33<-sum(Z1^2*dETA.ETA)

I44<-sum(Z2^2*dETA.ETA)

I55<-sum(D^2*dRHO1.RHO1)

I66<-sum((1-D)^2*dRHO1.RHO1)

I77<-sum(D^2*dRHO0.RHO0)

I88<-sum((1-D)^2*dRHO0.RHO0)

I99<-sum(Z1^2*(dRHO1.RHO1+dRHO0.RHO0+2*dRHO1.RHO0))

I1010<-sum(Z2^2*(dRHO1.RHO1+dRHO0.RHO0+2*dRHO1.RHO0))

I12<-sum(D*dETA.ETA)

I13<-sum(Z1*dETA.ETA)

I14<-sum(Z2*dETA.ETA)
```

```
I15<-sum(D*dETA.RHO1)

I16<-sum((1-D)*dETA.RHO1)

I17<-sum(D*dETA.RHO0)

I18<-sum((1-D)*dETA.RHO0)

I19<-sum(Z1*(dETA.RHO1+dETA.RHO0))

I110<-sum(Z2*(dETA.RHO1+dETA.RHO0))

I23<-sum(D*Z1*dETA.ETA)

I24<-sum(D*Z2*dETA.ETA)

I25<-sum(D^2*dETA.RHO1)

I26<-0

I27<-sum(D^2*dETA.RHO0)

I28<-0

I29<-sum(D*Z1*(dETA.RHO1+dETA.RHO0))

I210<-sum(D*Z2*(dETA.RHO1+dETA.RHO0))

I34<-sum(Z1*Z2*dETA.ETA)

I35<-sum(Z1*D*dETA.RHO1)

I36<-sum(Z1*(1-D)*dETA.RHO1)

I37<-sum(Z1*D*dETA.RHO0)

I38<-sum(Z1*(1-D)*dETA.RHO0)

I39<-sum(Z1^2*(dETA.RHO1+dETA.RHO0))

I310<-sum(Z1*Z2*(dETA.RHO1+dETA.RHO0))

I45<-sum(Z2*D*dETA.RHO1)

I46<-sum(Z2*(1-D)*dETA.RHO1)

I47<-sum(Z2*D*dETA.RHO0)

I48<-sum(Z2*(1-D)*dETA.RHO0)

I49<-sum(Z1*Z2*(dETA.RHO1+dETA.RHO0))

I410<-sum(Z2^2*(dETA.RHO1+dETA.RHO0))

I56<-0

I57<-sum(D^2*dRHO1.RHO0)

I58<-0

I59<-sum(Z1*D*(dRHO1.RHO1+dRHO1.RHO0))

I510<-sum(Z2*D*(dRHO1.RHO1+dRHO1.RHO0))

I67<-0

I68<-sum((1-D)^2*dRHO1.RHO0)

I69<-sum(Z1*(1-D)*(dRHO1.RHO1+dRHO1.RHO0))

I610<-sum(Z2*(1-D)*(dRHO1.RHO1+dRHO1.RHO0))

I78<-0
```

```
I79<-sum(Z1*D*(dRHO0.RHO0+dRHO1.RHO0))

I710<-sum(Z2*D*(dRHO0.RHO0+dRHO1.RHO0))

I89<-sum(Z1*(1-D)*(dRHO0.RHO0+dRHO1.RHO0))

I810<-sum(Z2*(1-D)*(dRHO0.RHO0+dRHO1.RHO0))

I910<-sum(Z1*Z2*(dRHO1.RHO1+dRHO0.RHO0+2*dRHO1.RHO0))

I<-rbind(c(I11,I12,I13,I14,I15,I16,I17,I18,I19,I110),
         c(I12,I22,I23,I24,I25,I26,I27,I28,I29,I210),
         c(I13,I23,I33,I34,I35,I36,I37,I38,I39,I310),
c(I14,I24,I34,I44,I45,I46,I47,I48,I49,I410),
c(I15,I25,I35,I45,I55,I56,I57,I58,I59,I510),
c(I16,I26,I36,I46,I56,I66,I67,I68,I69,I610),
c(I17,I27,I37,I47,I57,I67,I77,I78,I79,I710),
c(I18,I28,I38,I48,I58,I68,I78,I88,I89,I810),
c(I19,I29,I39,I49,I59,I69,I79,I89,I99,I910),
c(I110,I210,I310,I410,I510,I610,I710,I810,I910,I1010))

return(I)
}
```

Log Likelihood - 2S Approach

```
l2Sample<-function(W,v,D) {
etaD<-exp(W[1]+D[,1]*W[2]+D[,3]*W[3]+D[,4]*W[4])
rho1D<-exp(W[5]*D[,1]+W[6]*(1-D[,1])+D[,3]*W[9]+D[,4]*W[10])
rho0D<-exp(W[7]*D[,1]+W[8]*(1-D[,1])+D[,3]*W[9]+D[,4]*W[10])

eta<-exp(W[1]+v[,1]*W[2]+v[,3]*W[3]+v[,4]*W[4])
rho1<-exp(W[5]*v[,1]+W[6]*(1-v[,1])+v[,3]*W[9]+v[,4]*W[10])
rho0<-exp(W[7]*v[,1]+W[8]*(1-v[,1])+v[,3]*W[9]+v[,4]*W[10])

-(sum((1-D[,5])*log(1+etaD+rho1D+etaD*rho0D)+D[,5]*log(rho0D+etaD*rho1D+rho1D*rho0D+etaD*rho1D*rho0D)-
log(1+etaD)-log(1+rho1D)-log(1+rho0D))+sum(v[,2]*log(eta)+v[,2]*v[,5]*log(rho1)+
v[,5]*(1-v[,2])*log(rho0)-log(1+eta)-v[,2]*log(1+rho1)-(1-v[,2])*log(1+rho0)))

}
```

Log Likelihood - 3S Approach

```
l3Sample<-function(W,v,D) {
etaD<-exp(W[1]+D[,1]*W[2]+D[,3]*W[3]+D[,4]*W[4])
rho1D<-exp(W[5]*D[,1]+W[6]*(1-D[,1])+D[,3]*W[9]+D[,4]*W[10])
rho0D<-exp(W[7]*D[,1]+W[8]*(1-D[,1])+D[,3]*W[9]+D[,4]*W[10])

eta<-exp(W[1]+v[,1]*W[2]+v[,3]*W[3]+v[,4]*W[4])
rho1<-exp(W[5]*v[,1]+W[6]*(1-v[,1])+v[,3]*W[9]+v[,4]*W[10])
rho0<-exp(W[7]*v[,1]+W[8]*(1-v[,1])+v[,3]*W[9]+v[,4]*W[10])

-(sum((1-D[,5])*log(1+etaD+rho1D+etaD*rho0D)+D[,5]*log(rho0D+etaD*rho1D+rho1D*rho0D+etaD*rho1D*rho0D)-
log(1+etaD)-log(1+rho1D)-log(1+rho0D))+sum(v[,2]*v[,5]*(log(eta)+log(rho1))+
v[,5]*(1-v[,2])*log(rho0)+v[,5]*(1-v[,2])*log(1+rho1)+
v[,2]*v[,5]*log(1+rho0)-v[,5]*log(rho0+eta*rho1+rho1*rho0+eta*rho1*rho0)+
v[,2]*(1-v[,5])*log(eta)+(1-v[,2])*(1-v[,5])*log(1+rho1)+v[,2]*(1-v[,5])*log(1+rho0)-
(1-v[,5])*log(1+eta+rho1+eta*rho0)))}
```

Data Generation for Bootstrap Step of Sample Size Determination Algorithm

```
datagenBinary3bootCov<-function(n1,n0,alpha,beta,gamma1,gamma2,kappa11,kappa01,kappa10,kappa00,
tau1,tau2,dat) {

Z1<-sample(dat[,3],(n1+n0),replace=TRUE)
Z2<-sample(dat[,4],(n1+n0),replace=TRUE)

indZ.1<-sort(sample(1:(n1+n0),n1))
indZ.0<-setdiff(1:(n1+n0),indZ.1)
z1_1<-Z1[indZ.1]
z1_0<-Z1[indZ.0]

z2_1<-Z2[indZ.1]
z2_0<-Z2[indZ.0]

P1<-unpack(alpha+beta+gamma1*z1_1+gamma2*z2_1)
P0<-unpack(alpha+gamma1*z1_0+gamma2*z2_0)

Y1<-rbinom(n1,1,P1)
Y0<-rbinom(n0,1,P0)

Atil1<-rep(0,n1)
Atil0<-rep(0,n0)

Th11<-Y1*unpack(kappa11+tau1*z1_1+tau2*z2_1)
Th10<-(1-Y1)*unpack(kappa10+tau1*z1_1+tau2*z2_1)
Th01<-Y0*unpack(kappa01+tau1*z1_0+tau2*z2_0)
Th00<-(1-Y0)*unpack(kappa00+tau1*z1_0+tau2*z2_0)

Obs1_1<-rbinom(sum(Y1),1,Th11[which(Th11!=0)])
Obs1_0<-rbinom(n1-sum(Y1),1,Th10[which(Th10!=0)])

Atil1[which(Y1==1)]<-Obs1_1
Atil1[which(Y1==0)]<-Obs1_0

Obs0_1<-rbinom(sum(Y0),1,Th01[which(Th01!=0)])
Obs0_0<-rbinom(n0-sum(Y0),1,Th00[which(Th00!=0)])

Atil0[which(Y0==1)]<-Obs0_1
Atil0[which(Y0==0)]<-Obs0_0

D<-c(rep(1,n1),rep(0,n0))
Z1<-c(z1_1,z1_0)
Z2<-c(z2_1,z2_0)
A<-c(Y1,Y0)
Atilda<-c(Atil1,Atil0)

dataset<-cbind(D,A,Z1,Z2,Atilda)

return(dataset)
}
```

Sample Size Determination Algorithm

```
rm(list=ls(all=TRUE))
source("E:\\Thesis\\Ch. 2 Sample Size and Logistic\\Code\\FINAL\\CH4\\Binary
 Regression functions 12-29-2014.r")
#source("Binary regression functions")

N1<-5000
N0<-5000
N<-N1+N0

numsim<-100
parList<-c(-2,-1,1,-0.5,2.5,2,-2,-3,0.2,0.1)

alpha<-    parList[1]
beta<-     parList[2]
psi1<-     parList[3]
psi2<-     parList[4]
kappa11<-  parList[5]
kappa01<-  parList[6]
kappa10<-  parList[7]
kappa00<-  parList[8]
tau1<-     parList[9]
tau2<-     parList[10]
```

```
SeedS<-{}
p<-10
IOMC<-matrix(0,10,10)
IVMC<-matrix(0,10,10)
IV1MC<-matrix(0,10,10)
IV0MC<-matrix(0,10,10)


P_m<-0.4
NN1<-{}

for (i in 1:100) { #Generate MC estimate of #{Atilde = 1}
data<-datagenBinary(N1,N0,alpha,beta,psi1,psi2,kappa11,kappa01,kappa10,kappa00,tau1,tau2,.6)
NN1<-c(NN1,sum(data[,5]))
}
n1<-mean(NN1)

M<-P_m*N
M1<-round(((n1)/N)*M)
M0<-M-M1

#Generate an original EHR dataset
data<-datagenBinary(N1,N0,alpha,beta,psi1,psi2,kappa11,kappa01,kappa10,kappa00,tau1,tau2,.6)

#MC estimates of n_{d,a,\tilde{a},z1}
MC_1101<-{}
MC_0101<-{}

MC_1011<-{}
MC_0011<-{}

MC_1100<-{}
MC_0100<-{}

MC_1010<-{}
MC_0010<-{}

for (i in 1:numsim) {

tempseedi<-runif(1,1,10000000)
set.seed(tempseedi)
data<-datagenBinary3bootCov(N1,N0,alpha,beta,psi1,psi2,kappa11,kappa01,kappa10,kappa00,tau1,tau2,data)

MC_1101<-c(MC_1101,length(which(data[,1]==1&data[,2]==1&data[,3]==1&data[,5]==0)))
MC_0101<-c(MC_0101,length(which(data[,1]==0&data[,2]==1&data[,3]==1&data[,5]==0)))

MC_1011<-c(MC_1011,length(which(data[,1]==1&data[,2]==0&data[,3]==1&data[,5]==1)))
MC_0011<-c(MC_0011,length(which(data[,1]==0&data[,2]==0&data[,3]==1&data[,5]==1)))

MC_1100<-c(MC_1100,length(which(data[,1]==1&data[,2]==1&data[,3]==0&data[,5]==0)))
MC_0100<-c(MC_0100,length(which(data[,1]==0&data[,2]==1&data[,3]==0&data[,5]==0)))

MC_1010<-c(MC_1010,length(which(data[,1]==1&data[,2]==0&data[,3]==0&data[,5]==1)))
MC_0010<-c(MC_0010,length(which(data[,1]==0&data[,2]==0&data[,3]==0&data[,5]==1)))


tempseedV2<-runif(1,1,10000000)
set.seed(tempseedV2)
v2S<-Val2SampleBinary3(data,M)

tempseedV3<-runif(1,1,10000000)
set.seed(tempseedV3)
v3S<-Val3SampleBinary3(data,M1,M0)

IOMC<-IOMC+(IS(data,c(alpha,beta,psi1,psi2,kappa11,kappa01,kappa10,kappa00,tau1,tau2))/N)
IVMC<-IVMC+(IV(v2S$Validation,c(alpha,beta,psi1,psi2,kappa11,kappa01,kappa10,kappa00,tau1,tau2))/M)
IV1MC<-IV1MC+(IV.1(v3S$Validation,c(alpha,beta,psi1,psi2,kappa11,kappa01,kappa10,kappa00,tau1,tau2))/M1)
IV0MC<-IV0MC+(IV.0(v3S$Validation,c(alpha,beta,psi1,psi2,kappa11,kappa01,kappa10,kappa00,tau1,tau2))/M0)

SeedS<-rbind(SeedS,c(tempseedi,tempseedV2,tempseedV3))

}
```

```
Io<-I0MC/numsim
Iv<-IVMC/numsim
Iv1<-IV1MC/numsim
Iv0<-IV0MC/numsim

B<-seq(.01,1,.01)
b<-c(0,1,rep(0,8))

f2<-function(x,B) qnorm(1-.05/(2*p))^2*t(b)%*%solve(N*Io+x*Iv)%*%b-B^2

#error checker for singularity of matrices for low values of m
thresh<-1000
res2<-{}
for(i in 1:length(B)){
mfinal<-tryCatch(ceiling(uniroot(f2,c(1,N),B=B[i])$root),error=function(e) -1)
if(mfinal!=-1) {
k<-0
x0<-mfinal
while(k!=thresh){
k<-1
while(diff(sign(c(f2(x0+1,B[i]),f2(x0+k,B[i]))))==0 & k<thresh ) k<-k+1
if(k<thresh) x0<-x0+k
if(k>=thresh) mfinal<-x0
}
}
evalRoot<-mfinal
res2<-rbind(res2,c(evalRoot,B[i]))
}

P1.3<-seq(.1,.9,(.9-.1)/8)
f3<-function(x,B,pv) qnorm(1-.05/(2*p))^2*t(b)%*%solve(N*Io+x*(pv*Iv1+(1-pv)*Iv0))%*%b-B^2

res3<-{}

for(i in 1:length(B)) {
for(j in 1:length(P1.3)) {
mfinal<-tryCatch(ceiling(uniroot(f3,c(1,N),B=B[i],pv=P1.3[j])$root),error=function(e) -1)
if(mfinal!=-1) {
k<-0
x0<-mfinal
while(k!=thresh){
k<-1
while(diff(sign(c(f3(x0+1,B[i],P1.3[j]),f3(x0+k,B[i],P1.3[j]))))==0 & k<thresh ) k<-k+1
if(k<thresh) x0<-x0+k
if(k>=thresh) mfinal<-x0
}
}
evalMax<-mfinal
res3<-rbind(res3,c(evalMax,P1.3[j],B[i]))
}
}

table2<-SortByP1(res3)

FinalTable<-cbind(res2[,2],res2[,1],table2[which(table2[,2]==unique(table2[,2])[1]),1],
table2[which(table2[,2]==unique(table2[,2])[2]),1],
table2[which(table2[,2]==unique(table2[,2])[3]),1],
table2[which(table2[,2]==unique(table2[,2])[4]),1],
table2[which(table2[,2]==unique(table2[,2])[5]),1],
table2[which(table2[,2]==unique(table2[,2])[6]),1],
table2[which(table2[,2]==unique(table2[,2])[7]),1],
table2[which(table2[,2]==unique(table2[,2])[8]),1],
table2[which(table2[,2]==unique(table2[,2])[9]),1])

colnames(FinalTable)<-c("B","2 samp","P1 = 0.1","P1 = 0.2","P1 = 0.3","P1 = 0.4",
"P1 = 0.5","P1 = 0.6","P1 = 0.7","P1 = 0.8","P1 = 0.9")
rbind(FinalTable)

MCsizes<-c(mean(MC_1101),mean(MC_0101),mean(MC_1011),mean(MC_0011),mean(MC_1100),
mean(MC_0100),mean(MC_1010),mean(MC_0010))
MCsizes<-rbind(c("n_{1101}","n_{0101}","n_{1101}","n_{0011}","n_{1100}","n_{0100}",
"n_{1010}","n_{0010}"),MCsizes)
```

Main Simulation Code

```
#####################
#
#  Simulation for binary regression
#
#  Built around Beta being the paramter of interest
#
#   USES A PARAMETER LIST FROM AN EXCEL FILE

rm(list=ls(all=TRUE))

source("C:\\THESIS FINAL WORK SIMULATION RESULTS AND CODE\\
Binary\\R functions for Binary Regression with Misclassification\\
Binary Regression functions 5-14-2014 CURRENT.r")

m<-1000
N1<-5000
N0<-5000
N<-N1+N0
n<-N

numsim<-1000
V.THRESHOLD<-10

alpha<-    -2.197
beta<-     1.099
psi1<-     0
psi2<-     0
kappa11<-  2.197
kappa01<-  2.442
kappa10<-  -2.197
kappa00<-  2.442
tau1<-     0
tau2<-     0

parList<-c(alpha,beta,psi1,psi2,kappa11,kappa01,kappa10,kappa00,tau1,tau2)

b<-diag(rep(1,10))

##########
#
# Generate MC estimates of a \tilde{a} group sizes to avoid sampling problems

temp11<-{}
temp00<-{}
temp10<-{}
temp01<-{}

temp3.11<-{}
temp3.00<-{}
temp3.10<-{}
temp3.01<-{}
P1<-seq(.1,.9,.1)
for(i in 1:100) {

data<-datagenBinary(N1,N0,alpha,beta,psi1,psi2,kappa11,kappa01,kappa10,kappa00,tau1,tau2,.6)
v2S<-Val2SampleBinary(data,m)
v.2<-v2S$Validation

temp11<-c(temp11,length(intersect(which(v.2[which(v.2[,5]==1),2]==0),which(v.2[,1]==1))))
temp00<-c(temp00,length(intersect(which(v.2[which(v.2[,5]==0),2]==1),which(v.2[,1]==0))))
temp10<-c(temp10,length(intersect(which(v.2[which(v.2[,5]==1),2]==0),which(v.2[,1]==0))))
temp01<-c(temp01,length(intersect(which(v.2[which(v.2[,5]==0),2]==1),which(v.2[,1]==1))))

for(v in 1:length(P1)) {
print(v)

m1<-round(P1[v]*m)

m0<-m-m1

v3S<-tryCatch(Val3SampleBinary(data,m1,m0)$Validation, error=function(e) matrix(-1,m,5))
v.3<-v3S

temp3.11<-c(temp3.11,length(intersect(which(v.3[which(v.3[,5]==1),2]==0),which(v.3[,1]==1))))
temp3.00<-c(temp3.00,length(intersect(which(v.3[which(v.3[,5]==0),2]==1),which(v.3[,1]==0))))
temp3.10<-c(temp3.10,length(intersect(which(v.3[which(v.3[,5]==1),2]==0),which(v.3[,1]==0))))
temp3.01<-c(temp3.01,length(intersect(which(v.3[which(v.3[,5]==0),2]==1),which(v.3[,1]==1))))

}
}

temp11<-mean(temp11)
temp00<-mean(temp00)
temp10<-mean(temp10)
temp01<-mean(temp01)

temp3.11<-apply(matrix(temp3.11,9),1,mean)
temp3.00<-apply(matrix(temp3.00,9),1,mean)
temp3.10<-apply(matrix(temp3.10,9),1,mean)
temp3.01<-apply(matrix(temp3.00,9),1,mean)

#
#
#
#########



#####
#
# Initialize
#
NN1<-{}

#number of MC obser
MM11.2<-{}
MM00.2<-{}
MM10.2<-{}
MM01.2<-{}

MM11.3<-{}
MM00.3<-{}
MM10.3<-{}
MM01.3<-{}

res.2<-{}
res.3<-{}
Vars.2<-{}
Vars.3<-{}
res.til<-{}
SE.til<-{}

SeedS<-{}

#
####

#
#######

for (B in 1:numsim) {

DataSeed<-round(runif(1,1,1000000))
set.seed(DataSeed)
data<-datagenBinary(N1,N0,alpha,beta,psi1,psi2,kappa11,kappa01,kappa10,kappa00,tau1,tau2,.6)

######
#
# glm() estimates
#
GLM<-summary(glm(formula=data[,5]~data[,1]+data[,3]+data[,4],family=binomial("logit")))
mle_til<-GLM$coefficients[,1]
SE_til<-GLM$coefficients[,2]
```

```
res.til<-rbind(res.til,mle_til)
SE.til<-rbind(SE.til,SE_til)

#
#
#########

Val2Seed<-round(runif(1,1,1000000))
set.seed(Val2Seed)
v2S<-Val2SampleBinary(data,m)

NN1<-c(NN1,sum(data[,5]))

v.2<-v2S$Validation
vc.2<-v2S$ValComp

if(temp11>=V.THRESHOLD & temp00>=V.THRESHOLD&temp10>=V.THRESHOLD & temp01>=V.THRESHOLD) {

SeedS<-c(SeedS,Val2Seed)
MM11.2<-c(MM11.2,length(intersect(which(v.2[which(v.2[,5]==1),2]==0),which(v.2[,1]==1))))
MM00.2<-c(MM00.2,length(intersect(which(v.2[which(v.2[,5]==0),2]==1),which(v.2[,1]==0))))
MM10.2<-c(MM10.2,length(intersect(which(v.2[which(v.2[,5]==1),2]==0),which(v.2[,1]==0))))
MM01.2<-c(MM01.2,length(intersect(which(v.2[which(v.2[,5]==0),2]==1),which(v.2[,1]==1))))

######
#
# MLEs 2 sample
#
mle<-nlminb(c(0,0,0,0,0,0,0,0,0,0),l2Sample,v=v.2,D=data)$par

#
######

###############
#
# Cov Matrix and SE of Beta
#

CovMatrix<-tryCatch(solve(IO(data,mle)+IV(v.2,mle)),error=function(e) matrix(-1,10,10))
Vars.2<-rbind(Vars.2,diag(CovMatrix))

#
################

}

if(temp11<V.THRESHOLD | temp00<V.THRESHOLD|temp10<V.THRESHOLD | temp01<V.THRESHOLD) {
mle<-rep(10000,10)
Vars.2<-rbind(Vars.2,rep(10000,10))

MM11.2<-c(MM11.2,10000)
MM00.2<-c(MM00.2,10000)
MM10.2<-c(MM10.2,10000)
MM01.2<-c(MM01.2,10000)
SeedS<-c(SeedS,-1)

}

res.2<-rbind(res.2,mle)

for (v in 1:length(P1)) {

m1<-round(P1[v]*m)
m0<-m-m1

if(NN1[B]>m1&(N-NN1[B])>m0&temp3.11[v]>=V.THRESHOLD&temp3.00[v]>=V.THRESHOLD&
temp3.10[v]>=V.THRESHOLD&temp3.01[v]>=V.THRESHOLD)     {

Val3Seed<-round(runif(1,1,1000000))
set.seed(Val3Seed)
v3S<-Val3SampleBinary(data,m1,m0)

v.3<-v3S$Validation
vc.3<-v3S$ValComp

MM11.3<-c(MM11.3,length(intersect(which(v.3[which(v.3[,5]==1),2]==0),which(v.3[,1]==1))))
MM00.3<-c(MM00.3,length(intersect(which(v.3[which(v.3[,5]==0),2]==1),which(v.3[,1]==0))))
MM10.3<-c(MM10.3,length(intersect(which(v.3[which(v.3[,5]==1),2]==0),which(v.3[,1]==0))))
MM01.3<-c(MM01.3,length(intersect(which(v.3[which(v.3[,5]==0),2]==1),which(v.3[,1]==1))))

SeedS<-c(SeedS,Val3Seed)
mle3<-nlminb(c(0,0,0,0,0,0,0,0,0,0),l3Sample,v=v.3,D=data)$par
res.3<-rbind(res.3,mle3)

###############
#
# Cov Matrix and SE 3 sample
#

CovMatrix<-tryCatch(solve(I0(data,mle3)+IV.1(v.3,mle3)+IV.0(v.3,mle3)),error=function(e) e=matrix(-1,10,10))
Vars.3<-rbind(Vars.3,diag(CovMatrix))

#
################
}

if(NN1[B]<=m1|(N-NN1[B])<=m0|temp3.11[v]<V.THRESHOLD|temp3.00[v]<V.THRESHOLD|
temp3.10[v]<V.THRESHOLD|temp3.01[v]<V.THRESHOLD)  {
res.3<-rbind(res.3,rep(10000,10))
Vars.3<-rbind(Vars.3,rep(10000,10))
SeedS<-c(SeedS,-1)
MM11.3<-c(MM11.3,10000)
MM00.3<-c(MM00.3,10000)
MM10.3<-c(MM10.3,10000)
MM01.3<-c(MM01.3,10000)
}
}
}  #end B


TruePar<-c(alpha,beta,psi1,psi2,kappa11,kappa01,kappa10,kappa00,tau1,tau2)
SeedS<-matrix(SeedS,B,(length(P1)+1),byrow=TRUE)
colnames(SeedS)<-c("2 Samp", "P1=0.1","P1=0.2","P1=0.3","P1=0.4","P1=0.5","P1=0.6","P1=0.7","P1=0.8","P1=0.9")
sigDig<-4

##################
#
# CONFIDENCE INTERVAL RESULTS
#
#

CI2<-{}
if(res.2[1,1]!=10000) {
for(k in 1:10) {
CI2<-round(c(CI2,length(which(res.2[,k]-1.96*sqrt(Vars.2[,k])<rep(TruePar[k],numsim)&
res.2[,k]+1.96*sqrt(Vars.2[,k]) > rep(TruePar[k],numsim)))/numsim),sigDig)
}
}
if(res.2[1,1]==10000) CI2<-rep(10000,10)

CItil<-{}
for(k in 1:4)
CItil<-round(c(CItil,length(which(res.til[,k]-1.96*sqrt(SE.til[,k])<rep(TruePar[k],numsim)&
res.til[,k]+1.96*sqrt(SE.til[,k]) >rep(TruePar[k],numsim)))/numsim),sigDig)

#NOTE: the 3 samp results are in blocks (rows) of P1 values eg. P1:1, P1:2,...,P1:9, cols are params
CI3<-matrix(0,9,10)

for(j in 1:9) {
tempCI3<-{}
ind<-seq(j,(numsim*9),9)
if(res.3[ind[1],1]!=10000) {
for(k in 1:10) {
tempCI3<-round(c(tempCI3,length(which(res.3[ind,k]-1.96*sqrt(Vars.3[ind,k])<rep(TruePar[k],numsim)&
res.3[ind,k]+1.96*sqrt(Vars.3[ind,k]) > rep(TruePar[k],numsim)))/numsim),sigDig)
}
```

```
}
if(res.3[ind[1],1]==10000) tempCI3<-rep(10000,10)
CI3[j,]<-tempCI3
}

#10000 signifies no solution due to small sampling error
ResultsCI<-cbind(c(CItil,rep(10000,6)),CI2,t(CI3))
dimnames(ResultsCI)<-list(c("alpha","beta","psi1","psi2","kappa11","kappa01",
"kappa10","kappa00","tau1","tau2"),
c("CItil","2 Samp", "P1=0.1","P1=0.2","P1=0.3","P1=0.4","P1=0.5","P1=0.6","P1=0.7","P1=0.8","P1=0.9"))

#
# CI end
#
###########

###########
#
#
# POINT ESTIMATES
#

tempPE3<-{}
for(j in 1:9) {
ind<-seq(j,(numsim*9),9)
tempPE3<-rbind(tempPE3,apply(res.3[ind,],2,mean))
}
ResultsPOINTESTIMATES<-cbind(c(round(apply(res.til,2,mean),sigDig),rep(10000,6)),
round(apply(res.2,2,mean),sigDig),round(t(tempPE3),sigDig))
dimnames(ResultsPOINTESTIMATES)<-list(c("alpha","beta","psi1","psi2","kappa11",
"kappa01","kappa10","kappa00","tau1","tau2"),
c("CItil","2 Samp", "P1=0.1","P1=0.2","P1=0.3","P1=0.4","P1=0.5","P1=0.6","P1=0.7","P1=0.8","P1=0.9"))

#
#
#
#
###########

###########
#
#
# ESTIMATES OF VAR
#
# ind[ind2] double checks for problems inverting the cov matrix
# 10000 denotes sampling probs

tempSE3<-{}
tempMC3<-{}
for(j in 1:9) {
ind<-seq(j,(numsim*9),9)
ind2<-which(Vars.3[ind,]==-1|Vars.3[ind,]>10000,arr.ind=TRUE)
if(length(ind2)>0) {
tempSE3<-rbind(tempSE3,apply(Vars.3[ind[-ind2[,1]],],2,mean))
tempMC3<-rbind(tempMC3,apply(res.3[ind[-ind2[,1]],],2,var))
}
if(length(ind2)==0) {
tempSE3<-rbind(tempSE3,apply(Vars.3[ind,],2,mean))
tempMC3<-rbind(tempMC3,apply(res.3[ind,],2,var))
}
}

ResultsVARS<-cbind(c(round(apply(Vars.2,2,mean),sigDig)),round(t(tempSE3),sigDig))
ResultsVARS.MC<-cbind(c(round(apply(res.2,2,var),sigDig)),round(t(tempMC3),sigDig))
ResultsVARS.MC[which(ResultsVARS.MC==0)]<-10000
ResultsVARS[which(ResultsVARS==0)]<-10000
dimnames(ResultsVARS)<-list(c("alpha","beta","psi1","psi2","kappa11",
"kappa01","kappa10","kappa00","tau1","tau2"),
c("2 Samp", "P1=0.1","P1=0.2","P1=0.3","P1=0.4","P1=0.5","P1=0.6","P1=0.7","P1=0.8","P1=0.9"))
dimnames(ResultsVARS.MC)<-list(c("alpha","beta","psi1","psi2","kappa11","kappa01",
"kappa10","kappa00","tau1","tau2"),
c("2 Samp", "P1=0.1","P1=0.2","P1=0.3","P1=0.4","P1=0.5","P1=0.6","P1=0.7","P1=0.8","P1=0.9"))
```

```
#
#
#
#
###########
RESmm11<-{}
RESmm00<-{}
RESmm10<-{}
RESmm01<-{}

for(j in 1:9) {
ind<-seq(j,(numsim*9),9)
RESmm11<-cbind(RESmm11,MM11.3[ind])
RESmm00<-cbind(RESmm00,MM00.3[ind])
RESmm10<-cbind(RESmm10,MM10.3[ind])
RESmm01<-cbind(RESmm01,MM01.3[ind])
}

RESmm11<-cbind(MM11.2,RESmm11)
RESmm00<-cbind(MM00.2,RESmm00)
RESmm10<-cbind(MM10.2,RESmm10)
RESmm01<-cbind(MM01.2,RESmm01)

colnames(RESmm11)<-c("2 Samp", "P1=0.1","P1=0.2","P1=0.3","P1=0.4","P1=0.5",
"P1=0.6","P1=0.7","P1=0.8","P1=0.9")
colnames(RESmm00)<-c("2 Samp", "P1=0.1","P1=0.2","P1=0.3","P1=0.4","P1=0.5",
"P1=0.6","P1=0.7","P1=0.8","P1=0.9")
colnames(RESmm10)<-c("2 Samp", "P1=0.1","P1=0.2","P1=0.3","P1=0.4","P1=0.5",
"P1=0.6","P1=0.7","P1=0.8","P1=0.9")
colnames(RESmm01)<-c("2 Samp", "P1=0.1","P1=0.2","P1=0.3","P1=0.4","P1=0.5",
"P1=0.6","P1=0.7","P1=0.8","P1=0.9")
```

C.4 Chapter 5 Code

Data Generation $L^{(1)}$

```
datagenL1<-function(Th,Lambda,Beta,Psi1,Psi2,Mu,n,Pd=0.4,Pz1=0.6,m_prop) {
D<-rbinom(n,1,Pd)
Z1<-rbinom(n,1,Pz1)
Z2<-rgamma(n,2,1)

T<-rexp(n,Lambda*exp(Beta*D+Psi1*Z1+Psi2*Z2))
IndClass<-rbinom(n,1,Th)
C<-rexp(n,Mu)
Ttil<-pmin(T,C)*IndClass

Ttil[which(IndClass==0)]<-C[which(IndClass==0)]

deltaTil<-rep(0,n)

deltaTil[which(Ttil==T)]<-1
delta<-as.numeric(T<C)

dataset<-cbind(Ttil,T,C,deltaTil,delta,D,Z1,Z2)

m<-round(n*m_prop)

ValInd<-sample(which(deltaTil==0),m)

list(v=dataset[ValInd,],D=dataset)

}
```

Log-Likelihood $L^{(1)}$

```
ll.L1<-function(W,val,dat) {
#W = c(theta, lam,beta,psi1,psi2)

f_t<-log(W[2])+(W[3]*dat[,6]+W[4]*dat[,7]+W[5]*dat[,8])-W[2]*exp(W[3]*dat[,6]+
 W[4]*dat[,7]+W[5]*dat[,8])*dat[,1]
S_t<-     -W[2]*exp(W[3]*dat[,6]+W[4]*dat[,7]+W[5]*dat[,8])*dat[,1]
f_tV<-log(W[2])+(W[3]*val[,6]+W[4]*val[,7]+W[5]*val[,8])-W[2]*exp(W[3]*val[,6]+
W[4]*val[,7]+W[5]*val[,8])*val[,1]
S_tV<-     -W[2]*exp(W[3]*val[,6]+W[4]*val[,7]+W[5]*val[,8])*val[,1]

-(sum(dat[,4]*log(W[1])+dat[,4]*f_t+(1-dat[,4])*log((1-exp(S_t))*(1-W[1])+exp(S_t)))+
sum(val[,5]*(1-val[,4])*(log(1-W[1])+log(1-exp(S_tV)) - log(1-W[1]*(1-exp(S_tV)))  )+
(1-val[,5])*(1-val[,4])*(S_tV - log(1-W[1]*(1-exp(S_tV))))))}
```

Information Matrix - Original Sample $L^{(1)}$

```
IO<-function(data,param) {
#param = c(Th,Lambda,Beta,Psi1,Psi2)

t<-data[,1]
deltaTil<-data[,4]
delta<-data[,5]
d<-data[,6]
z1<-data[,7]
z2<-data[,8]

th<-param[1]
la<-param[2]
be<-param[3]
ps1<-param[4]
ps2<-param[5]

St<- exp(-t*la*exp(be*d+ps1*z1+ps2*z2))
ft<- la*exp(be*d+ps1*z1+ps2*z2)*St
```

```
DlDf<- deltaTil/ft
DlDs<- ((1-deltaTil)*th)/(1-th*(1-St))
D2lD2f<- -(deltaTil/ft^2)
D2lD2s<- -(((1-deltaTil)*th^2)/(1-th*(1-St))^2)
D2lDsDf<- 0

DfDla<- exp(be*d+ps1*z1+ps2*z2)*(St-t*ft)

DfDbe<- exp(be*d+ps1*z1+ps2*z2)*St*la*d*(1-t*la*exp(be*d+ps1*z1+ps2*z2))
DfDps1<- exp(be*d+ps1*z1+ps2*z2)*St*la*z1*(1-t*la*exp(be*d+ps1*z1+ps2*z2))
DfDps2<- exp(be*d+ps1*z1+ps2*z2)*St*la*z2*(1-t*la*exp(be*d+ps1*z1+ps2*z2))

DsDla<- -t*exp(be*d+ps1*z1+ps2*z2)*St
DsDbe<- -t*d*ft
DsDps1<- -t*z1*ft
DsDps2<- -t*z2*ft

D2fD2la<- t*exp(2*(be*d+ps1*z1+ps2*z2))*(t*ft-2*St)
D2fD2be<- d^2*ft*(1-3*la*t*exp(be*d+ps1*z1+ps2*z2)+exp(2*(be*d+ps1*z1+ps2*z2))*la^2*t^2)
D2fD2ps1<- z1^2*ft*(1-3*la*t*exp(be*d+ps1*z1+ps2*z2)+exp(2*(be*d+ps1*z1+ps2*z2))*la^2*t^2)
D2fD2ps2<- z2^2*ft*(1-3*la*t*exp(be*d+ps1*z1+ps2*z2)+exp(2*(be*d+ps1*z1+ps2*z2))*la^2*t^2)

D2sD2la<- t^2*exp(2*(be*d+ps1*z1+ps2*z2))*St
D2sD2be<- -t*d^2*ft*(1-la*t*exp(be*d+ps1*z1+ps2*z2))
D2sD2ps1<- -t*z1^2*ft*(1-la*t*exp(be*d+ps1*z1+ps2*z2))
D2sD2ps2<- -t*z2^2*ft*(1-la*t*exp(be*d+ps1*z1+ps2*z2))

D2sDlaDbe<- -t*d*exp(be*d+ps1*z1+ps2*z2)*St*(1-la*t*exp(be*d+ps1*z1+ps2*z2))
D2sDlaDps1<- -t*z1*exp(be*d+ps1*z1+ps2*z2)*St*(1-la*t*exp(be*d+ps1*z1+ps2*z2))
D2sDlaDps2<- -t*z2*exp(be*d+ps1*z1+ps2*z2)*St*(1-la*t*exp(be*d+ps1*z1+ps2*z2))
D2sDbeDps1<- -d*z1*t*ft*(1-la*t*exp(be*d+ps1*z1+ps2*z2))
D2sDbeDps2<- -d*z2*t*ft*(1-la*t*exp(be*d+ps1*z1+ps2*z2))
D2sDps1Dps2<- -z1*z2*t*ft*(1-la*t*exp(be*d+ps1*z1+ps2*z2))

D2fDlaDbe<- d*exp(be*d+ps1*z1+ps2*z2)*St*(1-3*la*t*exp(be*d+ps1*z1+ps2*z2)+
la^2*t^2*exp(2*(be*d+ps1*z1+ps2*z2)))
D2fDlaDps1<- z1*exp(be*d+ps1*z1+ps2*z2)*St*(1-3*la*t*exp(be*d+ps1*z1+ps2*z2)+
la^2*t^2*exp(2*(be*d+ps1*z1+ps2*z2)))
D2fDlaDps2<- z2*exp(be*d+ps1*z1+ps2*z2)*St*(1-3*la*t*exp(be*d+ps1*z1+ps2*z2)+
la^2*t^2*exp(2*(be*d+ps1*z1+ps2*z2)))
D2fDbeDps1<-  d*z1*la*exp(be*d+ps1*z1+ps2*z2)*St*(1-3*la*t*exp(be*d+ps1*z1+ps2*z2)+
la^2*t^2*exp(2*(be*d+ps1*z1+ps2*z2)))
D2fDbeDps2<-  d*z2*la*exp(be*d+ps1*z1+ps2*z2)*St*(1-3*la*t*exp(be*d+ps1*z1+ps2*z2)+
la^2*t^2*exp(2*(be*d+ps1*z1+ps2*z2)))
D2fDps1Dps2<- z1*z2*la*exp(be*d+ps1*z1+ps2*z2)*St*(1-3*la*t*exp(be*d+ps1*z1+ps2*z2)+
la^2*t^2*exp(2*(be*d+ps1*z1+ps2*z2)))


I11<- -sum(-deltaTil/th^2 - (1-deltaTil)*(1-St)^2/(1-th*(1-St))^2)

I22<- -sum(DfDla^2 * D2lD2f + DsDla^2 * D2lD2s + 2 * DfDla * DsDla * D2lDsDf +
DlDf * D2fD2la + DlDs * D2sD2la)

I33<- -sum(DfDbe^2 * D2lD2f + DsDbe^2 * D2lD2s + 2 * DfDbe * DsDbe * D2lDsDf +
DlDf * D2fD2be + DlDs * D2sD2be)

I44<- -sum(DfDps1^2 * D2lD2f + DsDps1^2 * D2lD2s + 2 * DfDps1 * DsDps1 * D2lDsDf +
DlDf * D2fD2ps1 + DlDs * D2sD2ps1)

I55<- -sum(DfDps2^2 * D2lD2f + DsDps2^2 * D2lD2s + 2 * DfDps2 * DsDps2 * D2lDsDf +
DlDf * D2fD2ps2 + DlDs * D2sD2ps2)

I12<- -sum((1-deltaTil)/(1-th * (1-St))^2 * DsDla)

I13<- -sum((1-deltaTil)/(1-th * (1-St))^2 * DsDbe)

I14<- -sum((1-deltaTil)/(1-th * (1-St))^2 * DsDps1)

I15<- -sum((1-deltaTil)/(1-th * (1-St))^2 * DsDps2)

I23<- -sum(DfDla * DfDbe * D2lD2f + DsDla * DsDbe * D2lD2s +
D2lDsDf * (DfDla * DsDbe + DfDbe * DsDla) + DlDf * D2fDlaDbe + DlDs * D2sDlaDbe)
```

```
I24<- -sum(DfDla * DfDps1 * D2lD2f + DsDla * DsDps1 * D2lD2s +
D2lDsDf * (DfDla * DsDps1 + DfDps1 * DsDla) + DlDf * D2fDlaDps1 + DlDs * D2sDlaDps1)

I25<- -sum(DfDla * DfDps2 * D2lD2f + DsDla * DsDps2 * D2lD2s +
D2lDsDf * (DfDla * DsDps2 + DfDps2 * DsDla) + DlDf * D2fDlaDps2 + DlDs * D2sDlaDps2)

I34<- -sum(DfDbe * DfDps1 * D2lD2f + DsDbe * DsDps1 * D2lD2s +
D2lDsDf * (DfDbe * DsDps1 + DfDps1 * DsDbe) + DlDf * D2fDbeDps1 + DlDs * D2sDbeDps1)

I35<- -sum(DfDbe * DfDps2 * D2lD2f + DsDbe * DsDps2 * D2lD2s +
D2lDsDf * (DfDbe * DsDps2 + DfDps2 * DsDbe) + DlDf * D2fDbeDps2 + DlDs * D2sDbeDps2)

I45<- -sum(DfDps1 * DfDps2 * D2lD2f + DsDps1 * DsDps2 * D2lD2s +
D2lDsDf * (DfDps1 * DsDps2 + DfDps2 * DsDps1) + DlDf * D2fDps1Dps2 + DlDs * D2sDps1Dps2)

I<-rbind(c(I11, I12, I13, I14, I15),c(I12, I22, I23, I24, I25),
c(I13, I23, I33, I34, I35),c(I14, I24, I34, I44, I45),
c(I15, I25, I35, I45, I55))

return(I)
}
```

Information Matrix - Validation Sample $L^{(1)}$

```
IV<-function(data,param) {
#param = c(Th,Lambda,Beta,Psi1,Psi2)

t<-data[,1]
deltaTil<-data[,4]
delta<-data[,5]
d<-data[,6]
z1<-data[,7]
z2<-data[,8]

th<-param[1]
la<-param[2]
be<-param[3]
ps1<-param[4]
ps2<-param[5]

St<- exp(-t*la*exp(be*d+ps1*z1+ps2*z2))
ft<- la*exp(be*d+ps1*z1+ps2*z2)*St

DlDs<- -((delta*(1-deltaTil))/(1-St))+((1-delta)*(1-deltaTil))/St-((1-deltaTil)*th)/(1-th*(1-St))
D2lD2s<- -((delta*(1-deltaTil))/(1-St)^2)-((1-delta)*(1-deltaTil))/St^2+((1-deltaTil)*th^2)/(1-th*(1-St))^2

DsDla<- -t*exp(be*d+ps1*z1+ps2*z2)*St
DsDbe<- -t*la*d*exp(be*d+ps1*z1+ps2*z2)*St
DsDps1<- -t*la*z1*exp(be*d+ps1*z1+ps2*z2)*St
DsDps2<- -t*la*z2*exp(be*d+ps1*z1+ps2*z2)*St

D2sD2la<- t^2*exp(2*(be*d+ps1*z1+ps2*z2))*St
D2sD2be<-  -d^2*ft*t*(1-la*t*exp(be*d+ps1*z1+ps2*z2))
D2sD2ps1<- -z1^2*ft*t*(1-la*t*exp(be*d+ps1*z1+ps2*z2))
D2sD2ps2<- -z2^2*ft*t*(1-la*t*exp(be*d+ps1*z1+ps2*z2))

D2sDlaDbe<-  -d*t*exp(be*d+ps1*z1+ps2*z2)*St*(1-la*t*exp(be*d+ps1*z1+ps2*z2))
D2sDlaDps1<-  -z1*t*exp(be*d+ps1*z1+ps2*z2)*St*(1-la*t*exp(be*d+ps1*z1+ps2*z2))
D2sDlaDps2<-  -z2*t*exp(be*d+ps1*z1+ps2*z2)*St*(1-la*t*exp(be*d+ps1*z1+ps2*z2))
D2sDbeDps1<-  -d*z1*t*ft*(1-la*t*exp(be*d+ps1*z1+ps2*z2))
D2sDbeDps2<-  -d*z2*t*ft*(1-la*t*exp(be*d+ps1*z1+ps2*z2))
D2sDps1Dps2<-  -z1*z2*t*ft*(1-la*t*exp(be*d+ps1*z1+ps2*z2))

I11<- -sum(-((delta*(1-deltaTil))/(1-th)^2)+((1-deltaTil)*(1-St)^2)/(1-th*(1-St))^2)

I22<- -sum(DsDla^2 * D2lD2s + DlDs * D2sD2la)

I33<- -sum(DsDbe^2 * D2lD2s + DlDs * D2sD2be)

I44<- -sum(DsDps1^2 * D2lD2s + DlDs * D2sD2ps1)
I55<- -sum(DsDps2^2 * D2lD2s + DlDs * D2sD2ps2)

I12<- -sum((((1-deltaTil)*exp(be*d+ps1*z1+ps2*z2)*St*t)/(1-(1-St)*th)^2))

I13<- -sum((((1-deltaTil)*exp(be*d+ps1*z1+ps2*z2)*St*t*la*d)/(1-(1-St)*th)^2))

I14<- -sum((((1-deltaTil)*exp(be*d+ps1*z1+ps2*z2)*St*t*la*z1)/(1-(1-St)*th)^2))

I15<- -sum((((1-deltaTil)*exp(be*d+ps1*z1+ps2*z2)*St*t*la*z2)/(1-(1-St)*th)^2))

I23<- -sum(DsDla * DsDbe * D2lD2s + DlDs * D2sDlaDbe)

I24<- -sum(DsDla * DsDps1 * D2lD2s + DlDs * D2sDlaDps1)

I25<- -sum(DsDla * DsDps2 * D2lD2s + DlDs * D2sDlaDps2)

I34<- -sum(DsDbe * DsDps1 * D2lD2s + DlDs * D2sDbeDps1)

I35<- -sum(DsDbe * DsDps2 * D2lD2s + DlDs * D2sDbeDps2)

I45<- -sum(DsDps1 * DsDps2 * D2lD2s + DlDs * D2sDps1Dps2)


I<-rbind(c(I11, I12, I13, I14, I15),c(I12, I22, I23, I24, I25),
c(I13, I23, I33, I34, I35),c(I14, I24, I34, I44, I45),
c(I15, I25, I35, I45, I55))

return(I)
}
```

Main Simulation Code

```
library(survival)
source("E:\\Thesis\\Ch. 3 Failure time\\FINAL CODE\\UNI2\\
failure time type and time Information Matrices 9-27-2014.r")

numsim<-1000
Theta<-0.99
Lam<-0.05
Beta<-0
Psi1<-0.4
Psi2<-0
Mu<-0.05
n<-10000
Pd<-0.4
Pz1<-0.6
mprop<-0.05
parList<-c(Theta,Lam,Beta,Psi1,Psi2,Mu,n,Pd,Pz1,mprop)

res<-{}
resC<-{}
Vars<-{}
resPerfClass<-{}
VarsPerfClass<-{}

SeedS<-runif(numsim,0,10000000)

for(i in 1:numsim) {

set.seed(SeedS[i])
dat<-datagenFailure.4par(parList)

d<-dat$D
val<-dat$v

mle<-nlminb(c(0.5,1,0,0,0),LL4,lower=c(0.000000001,0.000000001,-Inf,-Inf,-Inf),
upper=c(.99999999999,Inf,Inf,Inf,Inf),val=val,dat=d)$par
Vars<-rbind(Vars,diag(tryCatch(solve(IO(d,mle)+IV(val,mle)),error=function(e) e=matrix(10000,10,10))))
res<-rbind(res,mle)

PerfClassReg<-survreg(Surv(dat$D[,1],dat$D[,4]) ~ dat$D[,6]+dat$D[,7]+dat$D[,8],dist="exponential")
resPerfClass<-rbind(resPerfClass,c(1/exp(PerfClassReg$coefficients[1]),-PerfClassReg$coefficients[2:4]))
VarsPerfClass<-rbind(VarsPerfClass,c(PerfClassReg$var[1]/exp(2*PerfClassReg$coefficients[1]),
diag(PerfClassReg$var)[2:4]))

}

True<-parList
CI<-{}
for(l in 1:5)
CI<-c(CI,length(which(res[which(is.na(Vars[,l])==FALSE),l]-1.96*sqrt(Vars[which(is.na(Vars[,l])==FALSE),l])<
rep(True[l],length(which(is.na(Vars[,l])==FALSE))&
res[which(is.na(Vars[,l])==FALSE),l]+1.96*sqrt(Vars[which(is.na(Vars[,l])==FALSE),l]) >
rep(True[l],length(which(is.na(Vars[,l])==FALSE)))
))/(numsim-length(which(is.na(Vars[,l])==TRUE))))

numberNA<-length(which(is.na(Vars[,l])==TRUE))

CItil<-{}
for(l in 1:4)
CItil<-c(CItil,length(which(resPerfClass[,l]-1.96*sqrt(VarsPerfClass[,l])<rep(True[l+1],numsim)&
resPerfClass[,l]+1.96*sqrt(VarsPerfClass[,l]) >rep(True[l+1],numsim)))/numsim)

gammaTil<-apply(resPerfClass,2,mean)
gammaTilSE<-apply(VarsPerfClass,2,mean)

gammaHat<-apply(res,2,mean)[1:5]
gammaHatSE<-sqrt(apply(Vars[which(Vars[,1]!=10000),],2,mean))[1:5]

Results<-as.data.frame(cbind(as.character(True[1:5]),c("-",as.character(round(gammaTil,5))),
c("-",as.character(round(sqrt(gammaTilSE),5))),c("-",as.character(CItil)),
round(gammaHat,5),round(gammaHatSE,5),round(CI,3)))

colnames(Results)<-c("True","gamTil","gamTilSE","CItil","gamHat","gamHatSE","CI")
rownames(Results)<-c("Theta","Lambda","Beta","Psi1","Psi2")
}
```

Data Generation $L^{(2)}$

```
datagenL2<-function(Lambda1,Beta1,Psi1,Psi2,Lambda2,Beta2,Eta1,Eta2,Phi,Xi1, Xi2,
Mu,n,Pd,Pz1,paramG=c(2,1),m,pM1) {

D<-rbinom(n,1,Pd)
Z1<-rbinom(n,1,Pz1)
Z2<-rgamma(n,.5,1)

H1<-Lambda1*exp(Beta1*D+Psi1*Z1+Psi2*Z2)
H2<-Lambda2*exp(Beta2*D+Eta1*Z1+Eta2*Z2)

ACH<-H1+H2
T<-rexp(n,ACH)
C<-rexp(n,Mu)
J<-rbinom(n,1,H1/ACH)
J[which(J==0)]<-2

X<-pmin(T,C)

delta<-as.numeric(T<C)

deltaTil<-rep(0,n)
deltaTil[which(delta==1)]<-rbinom(length(which(delta==1)),1,Phi)
Xtil<-X

Xtil[which(deltaTil==1)]<-T[which(deltaTil==1)]
Xtil[which(deltaTil==0)]<-C[which(deltaTil==0)]

IndXi1<-rbinom(n,1,Xi1)
IndXi2<-rbinom(n,1,Xi2)

Jtil<-J
Jtil[which(J==2&deltaTil==1&IndXi2==0)]<-1
Jtil[which(J==1&deltaTil==1&IndXi1==0)]<-2
dataset<-cbind(Xtil,T,C,deltaTil,delta,Jtil,J,D,Z1,Z2)

M1<-round(m*pM1)
M0<-m-M1
ValInd1<-sample(1:length(which(dataset[,4]==1)),M1)
ValInd0<-sample(1:length(which(dataset[,4]==0)),M0)

Val<-rbind(dataset[which(dataset[,4]==1),][ValInd1,],dataset[which(dataset[,4]==0),][ValInd0,])

list(v=Val,D=dataset)
}
```

Log-Likelihood $L^{(2)}$

```
ll.L2<-function(W,dat,val) {
#W = c(Lambda1,Beta1,Psi1,Psi2,Lambda2,Beta2,Eta1,Eta2,Phi,Xi1,Xi2)
#Xtil,T,C,deltaTil,delta,Jtil,J,D,Z1,Z2

H1<-W[1]*exp(W[2]*dat[,8]+W[3]*dat[,9]+W[4]*dat[,10])
H2<-W[5]*exp(W[6]*dat[,8]+W[7]*dat[,9]+W[8]*dat[,10])
S_t<-     exp(-(H1+H2)*dat[,1])
f_t1<-H1*S_t
f_t2<-H2*S_t

H1V<-W[1]*exp(W[2]*val[,8]+W[3]*val[,9]+W[4]*val[,10])
H2V<-W[5]*exp(W[6]*val[,8]+W[7]*val[,9]+W[8]*val[,10])

S_tV<-     exp(-(H1V+H2V)*val[,1])
f_t1V<-H1V*S_tV
f_t2V<-H2V*S_tV

-(sum(dat[,4]*as.numeric(dat[,6]==1)*(log(W[9])-(H1+H2)*dat[,1]+log(W[10]*H1+(1-W[11])*H2))+
dat[,4]*as.numeric(dat[,6]==2)*(log(W[9])-(H1+H2)*dat[,1]+log((1-W[10])*H1+W[11]*H2))+
(1-dat[,4])*log(1-W[9]*(1-S_t)))+

sum((1-val[,4])*val[,5]*as.numeric(val[,7]==1)*log((1-W[9])*H1V/(H1V+H2V)*(1-S_tV)/(1-W[9]*(1-S_tV)))+
(1-val[,4])*val[,5]*as.numeric(val[,7]==2)*log((1-W[9])*H2V/(H1V+H2V)*(1-S_tV)/(1-W[9]*(1-S_tV)))-
(1-val[,4])*(1-val[,5])*((H1V+H2V)*val[,1]+log(1-W[9]*(1-S_tV))))+

sum(val[,4]*val[,5]*as.numeric(val[,6]==1)*as.numeric(val[,7]==1)*log(W[10]*H1V/(W[10]*H1V+(1-W[11])*H2V))+
val[,4]*val[,5]*as.numeric(val[,6]==1)*as.numeric(val[,7]==2)*log((1-W[11])*H2V/(W[10]*H1V+(1-W[11])*H2V))+
val[,4]*val[,5]*as.numeric(val[,6]==2)*as.numeric(val[,7]==1)*log((1-W[10])*H1V/((1-W[10])*H1V+W[11]*H2V))+
val[,4]*val[,5]*as.numeric(val[,6]==2)*as.numeric(val[,7]==2)*log(W[11]*H2V/((1-W[10])*H1V+W[11]*H2V))))
}
```

Information Matrix - Original Sample $L^{(2)}$

```
IO<-function(data,param) {
#data = Xtil,T,C,deltaTil,delta,Jtil,J,D,Z1,Z2
#param = c(Lambda1,Lambda2,Beta,alpha,Psi1,Psi2,Eta1,Eta2,Phi,Xi1,Xi2)

t<-data[,1]
deltaTil<-data[,4]
delta<-data[,5]
Jtil1<-as.numeric(data[,6]==1)
Jtil2<-as.numeric(data[,6]==2)
J1<-as.numeric(data[,7]==1)
J2<-as.numeric(data[,7]==2)
d<-data[,8]
z1<-data[,9]
z2<-data[,10]



la1<-param[1]
be1<-param[2]
ps1<-param[3]
ps2<-param[4]
la2<-param[5]
be2<-param[6]
```

```
et1<-param[7]
et2<-param[8]
ph<-param[9]
xi1<-param[10]
xi2<-param[11]


H1<-la1*exp(be1*d+ps1*z1+ps2*z2)
H2<-la2*exp(be2*d+et1*z1+et2*z2)

St<- exp(-t*(H1+H2))

#partial

DH1Dla1<-exp(be1*d+ps1*z1+ps2*z2)
DH1Dbe1<-d*la1*exp(be1*d+ps1*z1+ps2*z2)
DH1Dps1<-la1*z1*exp(be1*d+ps1*z1+ps2*z2)
DH1Dps2<-la1*z2*exp(be1*d+ps1*z1+ps2*z2)

DH2Dla2<-exp(be2*d+et1*z1+et2*z2)
DH2Dbe2<-la2*d*exp(be2*d+et1*z1+et2*z2)
DH2Det1<-la2*z1*exp(be2*d+et1*z1+et2*z2)
DH2Det2<-la2*z2*exp(be2*d+et1*z1+et2*z2)

DSDla1<--exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*t
DSDbe1<--d*la1*exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*t
DSDps1<--la1*z1*exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*t
DSDps2<--la1*z2*exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*t

DSDla2<--exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*t
DSDbe2<--d*la2*exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*t
DSDet1<--la2*z1*exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*t
DSDet2<--la2*z2*exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*t

D2H1Dla1la1<-0
D2H1Dla1be1<-d*exp(be1*d+ps1*z1+ps2*z2)
D2H1Dla1ps1<-z1*exp(be1*d+ps1*z1+ps2*z2)
D2H1Dla1ps2<-z2*exp(be1*d+ps1*z1+ps2*z2)
D2H1Dbe1be1<-la1*d^2*exp(be1*d+ps1*z1+ps2*z2)
D2H1Dbe1ps1<-la1*z1*d*exp(be1*d+ps1*z1+ps2*z2)
D2H1Dbe1ps2<-la1*z2*d*exp(be1*d+ps1*z1+ps2*z2)
D2H1Dps1ps1<-la1*z1^2*exp(be1*d+ps1*z1+ps2*z2)
D2H1Dps1ps2<-la1*z1*z2*exp(be1*d+ps1*z1+ps2*z2)
D2H1Dps2ps2<-la1*z2^2*exp(be1*d+ps1*z1+ps2*z2)

D2H2Dla2la2<-0
D2H2Dla2be2<-d*exp(be2*d+et1*z1+et2*z2)
D2H2Dla2et1<-z1*exp(be2*d+et1*z1+et2*z2)
D2H2Dla2et2<-z2*exp(be2*d+et1*z1+et2*z2)
D2H2Dbe2be2<-la2*d^2*exp(be2*d+et1*z1+et2*z2)
D2H2Dbe2et1<-la2*z1*d*exp(be2*d+et1*z1+et2*z2)
D2H2Dbe2et2<-la2*z2*d*exp(be2*d+et1*z1+et2*z2)
D2H2Det1et1<-la2*z1^2*exp(be2*d+et1*z1+et2*z2)
D2H2Det1et2<-la2*z1*z2*exp(be2*d+et1*z1+et2*z2)
D2H2Det2et2<-la2*z2^2*exp(be2*d+et1*z1+et2*z2)


D2SDla1la1<-exp(2*be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+2*ps1*z1+2*ps2*z2)*t^2
D2SDla1be1<--exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*t*d*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDla1ps1<--exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*t*z1*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDla1ps2<--exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*t*z2*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDla1la2<-exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDla1be2<-d*la2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDla1et1<-la2*z1*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDla1et2<-la2*z2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDbe1be1<--d*exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*la1*t*d*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDbe1ps1<--d*exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*la1*t*z1*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDbe1ps2<--d*exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*la1*t*z2*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDbe1la2<-d*la1*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDbe1be2<-d^2*la1*la2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDbe1et1<-d*la1*la2*z1*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDbe1et2<-d*la1*la2*z2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDps1ps1<--exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*la1*t*z1*z1*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDps1ps2<--exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*la1*t*z1*z2*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDps1la2<-la1*z1*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDps1be2<-d*la1*la2*z1*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDps1et1<-la1*la2*z1^2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDps1et2<-la1*la2*z1*z2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDps2ps2<--exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*la1*t*z2*z2*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDps2la2<-la1*z2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDps2be2<-d*la1*la2*z2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDps2et1<-la1*la2*z1*z2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDps2et2<-la1*la2*z2^2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDla2la2<-exp(2*be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+2*et1*z1+2*et2*z2)*t^2
D2SDla2be2<--exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*t*d*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)
D2SDla2et1<--exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*t*z1*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)
D2SDla2et2<--exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*t*z2*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)
D2SDbe2be2<--d*exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*la2*t*d*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)
D2SDbe2et1<--d*exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*la2*t*z1*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)
D2SDbe2et2<--d*exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*la2*t*z2*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)
D2SDet1et1<--exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*la2*t*z1*z1*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)
D2SDet1et2<--exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*la2*t*z1*z2*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)
D2SDet2et2<--exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*la2*t*z2*z2*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)

DH1Dla2<-0
DH1Dbe2<-0
DH1Det1<-0
DH1Det2<-0

DH2Dla1<-0
DH2Dbe1<-0
DH2Dps1<-0
DH2Dps2<-0

D2H2Dla1la2<-0
D2H2Dla1be2<-0
D2H2Dla1et1<-0
D2H2Dla1et2<-0
D2H2Dbe1la2<-0
D2H2Dbe1be2<-0
D2H2Dbe1et1<-0
D2H2Dbe1et2<-0
```

```
D2H2Dps1la2<-0
D2H2Dps1be2<-0
D2H2Dps1et1<-0
D2H2Dps1et2<-0
D2H2Dps2la2<-0
D2H2Dps2be2<-0
D2H2Dps2et1<-0
D2H2Dps2et2<-0

D2H1Dla2la1<-0
D2H1Dla2be1<-0
D2H1Dla2ps1<-0
D2H1Dla2ps2<-0
D2H1Dbe2la1<-0
D2H1Dbe2be1<-0
D2H1Dbe2ps1<-0
D2H1Dbe2ps2<-0
D2H1Det1la1<-0
D2H1Det1be1<-0
D2H1Det1ps1<-0
D2H1Det1ps2<-0
D2H1Det2la1<-0
D2H1Det2be1<-0
D2H1Det2ps1<-0
D2H1Det2ps2<-0

DlDph<-(deltaTil*(Jtil1+Jtil2))/ph-((1-deltaTil)*(1-St))/(1-ph*(1-St))
DlDxi1<-(deltaTil*H1*Jtil1)/(H1*xi1+H2*(1-xi2))-(deltaTil*H1*Jtil2)/(H1*(1-xi1)+H2*xi2)
DlDxi2<--((deltaTil*H2*Jtil1)/(H1*xi1+H2*(1-xi2)))+(deltaTil*H2*Jtil2)/(H1*(1-xi1)+H2*xi2)
DlDH1<-deltaTil*Jtil1*(xi1/(H1*xi1+H2*(1-xi2))-t)+deltaTil*Jtil2*((1-xi1)/(H1*(1-xi1)+H2*xi2)-t)
DlDH2<-deltaTil*Jtil1*((1-xi2)/(H1*xi1+H2*(1-xi2))-t)+deltaTil*Jtil2*(xi2/(H1*(1-xi1)+H2*xi2)-t)
DlDS<-((1-deltaTil)*ph)/(1-ph*(1-St))

D2lDphph<--((deltaTil*(Jtil1+Jtil2))/ph^2)-((1-deltaTil)*(1-St)^2)/(1-ph*(1-St))^2
D2lDphxi1<-0
D2lDphxi2<-0
D2lDphH1<-0
D2lDphH2<-0
D2lDphS<-(1-deltaTil)/(1-ph*(1-St))^2

D2lDxi1xi1<--((deltaTil*H1^2*Jtil1)/(H1*xi1+H2*(1-xi2))^2)-(deltaTil*H1^2*Jtil2)/(H1*(1-xi1)+H2*xi2)^2
D2lDxi1xi2<-(deltaTil*H1*H2*Jtil1)/(H1*xi1+H2*(1-xi2))^2+(deltaTil*H1*H2*Jtil2)/(H1*(1-xi1)+H2*xi2)^2
D2lDxi1H1<-((deltaTil*H2*Jtil1*(1-xi2))/(H1*xi1+H2*(1-xi2))^2)-(deltaTil*H2*Jtil2*xi2)/(H1*(1-xi1)+H2*xi2)^2
D2lDxi1H2<--((deltaTil*H1*Jtil1*(1-xi2))/(H1*xi1+H2*(1-xi2))^2)+(deltaTil*H1*Jtil2*xi2)/(H1*(1-xi1)+H2*xi2)^2
D2lDxi1S<-0

D2lDxi2xi2<--((deltaTil*H2^2*Jtil1)/(H1*xi1+H2*(1-xi2))^2)-(deltaTil*H2^2*Jtil2)/(H1*(1-xi1)+H2*xi2)^2
D2lDxi2H1<-(deltaTil*H2*Jtil1*xi1)/(H1*xi1+H2*(1-xi2))^2-(deltaTil*H2*Jtil2*(1-xi1))/(H1*(1-xi1)+H2*xi2)^2
D2lDxi2H2<-((deltaTil*H1*Jtil2*(1-xi1))/(H1*(1-xi1)+H2*xi2)^2)-(deltaTil*H1*Jtil1*xi1)/(H1*xi1+H2*(1-xi2))^2
D2lDxi2S<-0

D2lDH1H1<--((deltaTil*Jtil1*xi1^2)/(H1*xi1+H2*(1-xi2))^2)-(deltaTil*Jtil2*(1-xi1)^2)/(H1*(1-xi1)+H2*xi2)^2
D2lDH1H2<--((deltaTil*Jtil1*xi1*(1-xi2))/(H1*xi1+H2*(1-xi2))^2)-
(deltaTil*Jtil2*(1-xi1)*xi2)/(H1*(1-xi1)+H2*xi2)^2
D2lDH1S<-0

D2lDH2H2<--((deltaTil*Jtil1*(1-xi2)^2)/(H1*xi1+H2*(1-xi2))^2)-(deltaTil*Jtil2*xi2^2)/(H1*(1-xi1)+H2*xi2)^2
D2lDH2S<-0

D2lDSS<--(((1-deltaTil)*ph^2)/(1-ph*(1-St))^2)

D2lDH1ph<-0
D2lDH1xi1<-(deltaTil*Jtil1*H2*(1-xi2))/(H1*xi1+H2*(1- xi2))^2-(deltaTil*H2*Jtil2*xi2)/(H1*(1-xi1)+H2*xi2)^2
D2lDH1xi2<-(deltaTil*H2*Jtil1*xi1)/(H1*xi1+H2*(1-xi2))^2-(deltaTil*H2*Jtil2*(1-xi1))/(H1*(1-xi1)+H2*xi2)^2
D2lDH2ph<-0
D2lDH2xi1<-(deltaTil*H1*Jtil2*xi2)/(H1*(1-xi1)+H2*xi2)^2-((deltaTil*H1*Jtil1*(1-xi2))/(H1*xi1+H2*(1-xi2))^2)
D2lDH2xi2<-(deltaTil*H1*Jtil2*(1-xi1))/(H1*(1-xi1)+H2*xi2)^2-(deltaTil*H1*Jtil1*xi1)/(H1*xi1+H2*(1- xi2))^2
D2lDSph<-(1-deltaTil)/(1-ph*(1-St))^2
D2lDSxi1<-0
D2lDSxi2<-0
```

```
I11<--sum(DlDH1*D2H1Dla1la1+DlDS*D2SDla1la1+D2lDH1H1*(DH1Dla1)^2+D2lDSS*(DSDla1)^2+2*D2lDH1S*DH1Dla1*DSDla1)
I22<--sum(DlDH1*D2H1Dbe1be1+DlDS*D2SDbe1be1+D2lDH1H1*(DH1Dbe1)^2+D2lDSS*(DSDbe1)^2+2*D2lDH1S*DH1Dbe1*DSDbe1)
I33<--sum(DlDH1*D2H1Dps1ps1+DlDS*D2SDps1ps1+D2lDH1H1*(DH1Dps1)^2+D2lDSS*(DSDps1)^2+2*D2lDH1S*DH1Dps1*DSDps1)
I44<--sum(DlDH1*D2H1Dps2ps2+DlDS*D2SDps2ps2+D2lDH1H1*(DH1Dps2)^2+D2lDSS*(DSDps2)^2+2*D2lDH1S*DH1Dps2*DSDps2)
I55<--sum(DlDH2*D2H2Dla2la2+DlDS*D2SDla2la2+D2lDH2H2*(DH2Dla2)^2+D2lDSS*(DSDla2)^2+2*D2lDH2S*DH2Dla2*DSDla2)
I66<--sum(DlDH2*D2H2Dbe2be2+DlDS*D2SDbe2be2+D2lDH2H2*(DH2Dbe2)^2+D2lDSS*(DSDbe2)^2+2*D2lDH2S*DH2Dbe2*DSDbe2)
I77<--sum(DlDH2*D2H2Det1et1+DlDS*D2SDet1et1+D2lDH2H2*(DH2Det1)^2+D2lDSS*(DSDet1)^2+2*D2lDH2S*DH2Det1*DSDet1)
I88<--sum(DlDH2*D2H2Det2et2+DlDS*D2SDet2et2+D2lDH2H2*(DH2Det2)^2+D2lDSS*(DSDet2)^2+2*D2lDH2S*DH2Det2*DSDet2)
I99<--sum(D2lDphph)
I1010<--sum(D2lDxi1xi1)
I1111<--sum(D2lDxi2xi2)

I12<--sum(DlDH1*D2H1Dla1be1+DH1Dla1*(D2lDH1H1*DH1Dbe1+D2lDH1S*DSDbe1)+
DSDla1*(D2lDSS*DSDbe1+D2lDH1S*DH1Dbe1)+D2SDla1be1*DlDS)
I13<--sum(DlDH1*D2H1Dla1ps1+DH1Dla1*(D2lDH1H1*DH1Dps1+D2lDH1S*DSDps1)+
DSDla1*(D2lDSS*DSDps1+D2lDH1S*DH1Dps1)+D2SDla1ps1*DlDS)
I14<--sum(DlDH1*D2H1Dla1ps2+DH1Dla1*(D2lDH1H1*DH1Dps2+D2lDH1S*DSDps2)+
DSDla1*(D2lDSS*DSDps2+D2lDH1S*DH1Dps2)+D2SDla1ps2*DlDS)
I15<--sum(DlDH2*D2H2Dla1la2+DH2Dla2*(D2lDH2H2*DH2Dla1+D2lDH2S*DSDla1+
D2lDH1H2*DH1Dla1)+DSDla2*(D2lDSS*DSDla1+D2lDH2S*DH2Dla1+D2lDH1S*DH1Dla1)+D2SDla1la2*DlDS)
I16<--sum(DlDH2*D2H2Dla1be2+DH2Dbe2*(D2lDH2H2*DH2Dla1+D2lDH2S*DSDla1+
D2lDH1H2*DH1Dla1)+DSDbe2*(D2lDSS*DSDla1+D2lDH2S*DH2Dla1+D2lDH1S*DH1Dla1)+D2SDla1be2*DlDS)
I17<--sum(DlDH2*D2H2Dla1et1+DH2Det1*(D2lDH2H2*DH2Dla1+D2lDH2S*DSDla1+
D2lDH1H2*DH1Dla1)+DSDet1*(D2lDSS*DSDla1+D2lDH2S*DH2Dla1+D2lDH1S*DH1Dla1)+D2SDla1et1*DlDS)
I18<--sum(DlDH2*D2H2Dla1et2+DH2Det2*(D2lDH2H2*DH2Dla1+D2lDH2S*DSDla1+
D2lDH1H2*DH1Dla1)+DSDet2*(D2lDSS*DSDla1+D2lDH2S*DH2Dla1+D2lDH1S*DH1Dla1)+D2SDla1et2*DlDS)
I19<--sum(D2lDH1ph*DH1Dla1+D2lDSph*DSDla1)
I110<--sum(D2lDH1xi1*DH1Dla1+D2lDSxi1*DSDla1)
I111<--sum(D2lDH1xi2*DH1Dla1+D2lDSxi2*DSDla1)

I23<--sum(DlDH1*D2H1Dbe1ps1+DH1Dbe1*(D2lDH1H1*DH1Dps1+D2lDH1S*DSDps1)+
DSDbe1*(D2lDSS*DSDps1+D2lDH1S*DH1Dps1)+D2SDbe1ps1*DlDS)
I24<--sum(DlDH1*D2H1Dbe1ps2+DH1Dbe1*(D2lDH1H1*DH1Dps2+D2lDH1S*DSDps2)+
DSDbe1*(D2lDSS*DSDps2+D2lDH1S*DH1Dps2)+D2SDbe1ps2*DlDS)
I25<--sum(DlDH2*D2H2Dbe1la2+DH2Dla2*(D2lDH2H2*DH2Dbe1+D2lDH2S*DSDbe1+
D2lDH1H2*DH1Dbe1)+DSDla2*(D2lDSS*DSDbe1+D2lDH2S*DH2Dbe1+D2lDH1S*DH1Dbe1)+D2SDbe1la2*DlDS)
I26<--sum(DlDH2*D2H2Dbe1be2+DH2Dbe2*(D2lDH2H2*DH2Dbe1+D2lDH2S*DSDbe1+
D2lDH1H2*DH1Dbe1)+DSDbe2*(D2lDSS*DSDbe1+D2lDH2S*DH2Dbe1+D2lDH1S*DH1Dbe1)+D2SDbe1be2*DlDS)
I27<--sum(DlDH2*D2H2Dbe1et1+DH2Det1*(D2lDH2H2*DH2Dbe1+D2lDH2S*DSDbe1+
D2lDH1H2*DH1Dbe1)+DSDet1*(D2lDSS*DSDbe1+D2lDH2S*DH2Dbe1+D2lDH1S*DH1Dbe1)+D2SDbe1et1*DlDS)
I28<--sum(DlDH2*D2H2Dbe1et2+DH2Det2*(D2lDH2H2*DH2Dbe1+D2lDH2S*DSDbe1+
D2lDH1H2*DH1Dbe1)+DSDet2*(D2lDSS*DSDbe1+D2lDH2S*DH2Dbe1+D2lDH1S*DH1Dbe1)+D2SDbe1et2*DlDS)
I29<--sum(D2lDH1ph*DH1Dbe1+D2lDSph*DSDbe1)
I210<--sum(D2lDH1xi1*DH1Dbe1+D2lDSxi1*DSDbe1)
I211<--sum(D2lDH1xi2*DH1Dbe1+D2lDSxi2*DSDbe1)

I34<--sum(DlDH1*D2H1Dps1ps2+DH1Dps1*(D2lDH1H1*DH1Dps2+D2lDH1S*DSDps2)+
DSDps1*(D2lDSS*DSDps2+D2lDH1S*DH1Dps2)+D2SDps1ps2*DlDS)
I35<--sum(DlDH2*D2H2Dps1la2+DH2Dla2*(D2lDH2H2*DH2Dps1+D2lDH2S*DSDps1+
D2lDH1H2*DH1Dps1)+DSDla2*(D2lDSS*DSDps1+D2lDH2S*DH2Dps1+D2lDH1S*DH1Dps1)+D2SDps1la2*DlDS)
I36<--sum(DlDH2*D2H2Dps1be2+DH2Dbe2*(D2lDH2H2*DH2Dps1+D2lDH2S*DSDps1+
D2lDH1H2*DH1Dps1)+DSDbe2*(D2lDSS*DSDps1+D2lDH2S*DH2Dps1+D2lDH1S*DH1Dps1)+D2SDps1be2*DlDS)
I37<--sum(DlDH2*D2H2Dps1et1+DH2Det1*(D2lDH2H2*DH2Dps1+D2lDH2S*DSDps1+
D2lDH1H2*DH1Dps1)+DSDet1*(D2lDSS*DSDps1+D2lDH2S*DH2Dps1+D2lDH1S*DH1Dps1)+D2SDps1et1*DlDS)
I38<--sum(DlDH2*D2H2Dps1et2+DH2Det2*(D2lDH2H2*DH2Dps1+D2lDH2S*DSDps1+
D2lDH1H2*DH1Dps1)+DSDet2*(D2lDSS*DSDps1+D2lDH2S*DH2Dps1+D2lDH1S*DH1Dps1)+D2SDps1et2*DlDS)
I39<--sum(D2lDH1ph*DH1Dps1+D2lDSph*DSDps1)
I310<--sum(D2lDH1xi1*DH1Dps1+D2lDSxi1*DSDps1)
I311<--sum(D2lDH1xi2*DH1Dps1+D2lDSxi2*DSDps1)

I45<--sum(DlDH2*D2H2Dps2la2+DH2Dla2*(D2lDH2H2*DH2Dps2+D2lDH2S*DSDps2+D2lDH1H2*DH1Dps2)+
DSDla2*(D2lDSS*DSDps2+D2lDH2S*DH2Dps2+D2lDH1S*DH1Dps2)+D2SDps2la2*DlDS)
I46<--sum(DlDH2*D2H2Dps2be2+DH2Dbe2*(D2lDH2H2*DH2Dps2+D2lDH2S*DSDps2+D2lDH1H2*DH1Dps2)+
DSDbe2*(D2lDSS*DSDps2+D2lDH2S*DH2Dps2+D2lDH1S*DH1Dps2)+D2SDps2be2*DlDS)
I47<--sum(DlDH2*D2H2Dps2et1+DH2Det1*(D2lDH2H2*DH2Dps2+D2lDH2S*DSDps2+D2lDH1H2*DH1Dps2)+
DSDet1*(D2lDSS*DSDps2+D2lDH2S*DH2Dps2+D2lDH1S*DH1Dps2)+D2SDps2et1*DlDS)
I48<--sum(DlDH2*D2H2Dps2et2+DH2Det2*(D2lDH2H2*DH2Dps2+D2lDH2S*DSDps2+D2lDH1H2*DH1Dps2)+
DSDet2*(D2lDSS*DSDps2+D2lDH2S*DH2Dps2+D2lDH1S*DH1Dps2)+D2SDps2et2*DlDS)
I49<--sum(D2lDH1ph*DH1Dps2+D2lDSph*DSDps2)
I410<--sum(D2lDH1xi1*DH1Dps2+D2lDSxi1*DSDps2)
I411<--sum(D2lDH1xi2*DH1Dps2+D2lDSxi2*DSDps2)

I56<--sum(DlDH2*D2H2Dla2be2+DH2Dla2*(D2lDH2H2*DH2Dbe2+D2lDH2S*DSDbe2)+
DSDla2*(D2lDSS*DSDbe2+D2lDH2S*DH2Dbe2)+D2SDla2be2*DlDS)
I57<--sum(DlDH2*D2H2Dla2et1+DH2Dla2*(D2lDH2H2*DH2Det1+D2lDH2S*DSDet1)+
DSDla2*(D2lDSS*DSDet1+D2lDH2S*DH2Det1)+D2SDla2et1*DlDS)
I58<--sum(DlDH2*D2H2Dla2et2+DH2Dla2*(D2lDH2H2*DH2Det2+D2lDH2S*DSDet2)+
DSDla2*(D2lDSS*DSDet2+D2lDH2S*DH2Det2)+D2SDla2et2*DlDS)
I59<--sum(D2lDH2ph*DH2Dla2+D2lDSph*DSDla2)
I510<--sum(D2lDH2xi1*DH2Dla2+D2lDSxi1*DSDla2)
I511<--sum(D2lDH2xi2*DH2Dla2+D2lDSxi2*DSDla2)

I67<--sum(DlDH2*D2H2Dbe2et1+DH2Dbe2*(D2lDH2H2*DH2Det1+D2lDH2S*DSDet1)+
DSDbe2*(D2lDSS*DSDet1+D2lDH2S*DH2Det1)+D2SDbe2et1*DlDS)
I68<--sum(DlDH2*D2H2Dbe2et2+DH2Dbe2*(D2lDH2H2*DH2Det2+D2lDH2S*DSDet2)+
DSDbe2*(D2lDSS*DSDet2+D2lDH2S*DH2Det2)+D2SDbe2et2*DlDS)
I69<--sum(D2lDH2ph*DH2Dbe2+D2lDSph*DSDbe2)
I610<--sum(D2lDH2xi1*DH2Dbe2+D2lDSxi1*DSDbe2)
I611<--sum(D2lDH2xi2*DH2Dbe2+D2lDSxi2*DSDbe2)

I78<--sum(DlDH2*D2H2Det1et2+DH2Det1*(D2lDH2H2*DH2Det2+D2lDH2S*DSDet2)+
DSDet1*(D2lDSS*DSDet2+D2lDH2S*DH2Det2)+D2SDet1et2*DlDS)
I79<--sum(D2lDH2ph*DH2Det1+D2lDSph*DSDet1)
I710<--sum(D2lDH2xi1*DH2Det1+D2lDSxi1*DSDet1)
I711<--sum(D2lDH2xi2*DH2Det1+D2lDSxi2*DSDet1)

I89<--sum(D2lDH2ph*DH2Det2+D2lDSph*DSDet2)
I810<--sum(D2lDH2xi1*DH2Det2+D2lDSxi1*DSDet2)
I811<--sum(D2lDH2xi2*DH2Det2+D2lDSxi2*DSDet2)

I910<--sum(D2lDphxi1)
I911<--sum(D2lDphxi2)

I1011<--sum(D2lDxi1xi2)

I<-rbind(c(I11,I12,I13,I14,I15,I16,I17,I18,I19,I110,I111),
         c(I12,I22,I23,I24,I25,I26,I27,I28,I29,I210,I211),
         c(I13,I23,I33,I34,I35,I36,I37,I38,I39,I310,I311),
         c(I14,I24,I34,I44,I45,I46,I47,I48,I49,I410,I411),
         c(I15,I25,I35,I45,I55,I56,I57,I58,I59,I510,I511),
         c(I16,I26,I36,I46,I56,I66,I67,I68,I69,I610,I611),
         c(I17,I27,I37,I47,I57,I67,I77,I78,I79,I710,I711),
         c(I18,I28,I38,I48,I58,I68,I78,I88,I89,I810,I811),
         c(I19,I29,I39,I49,I59,I69,I79,I89,I99,I910,I911),
         c(I110,I210,I310,I410,I510,I610,I710,I810,I910,I1010,I1011),
         c(I111,I211,I311,I411,I511,I611,I711,I811,I911,I1011,I1111))


return(I)
}
```

Information Matrix - Validation Sample $L^{(2)}$

```
IV<-function(data,param) {
#data = Xtil,T,C,deltaTil,delta,D,Z1,Z2
#param = c(Th,Lambda,Beta,Psi1,Psi2)

#data = Xtil,T,C,deltaTil,delta,Jtil,J,D,Z1,Z2
#param = c(Phi,Xi1,Xi2,Lambda1,Lambda2,Beta,alpha,Psi1,Psi2,Eta1,Eta2)

t<-data[,1]
deltaTil<-data[,4]
delta<-data[,5]
Jtil1<-as.numeric(data[,6]==1)
Jtil2<-as.numeric(data[,6]==2)
J1<-as.numeric(data[,7]==1)
J2<-as.numeric(data[,7]==2)
d<-data[,8]
z1<-data[,9]
z2<-data[,10]



la1<-param[1]
```

```
be1<-param[2]
ps1<-param[3]
ps2<-param[4]
la2<-param[5]
be2<-param[6]
et1<-param[7]
et2<-param[8]
ph<-param[9]
xi1<-param[10]
xi2<-param[11]


H1<-la1*exp(be1*d+ps1*z1+ps2*z2)
H2<-la2*exp(be2*d+et1*z1+et2*z2)

St<- exp(-t*(H1+H2))

#partial

DH1Dla1<-exp(be1*d+ps1*z1+ps2*z2)
DH1Dbe1<-d*la1*exp(be1*d+ps1*z1+ps2*z2)
DH1Dps1<-la1*z1*exp(be1*d+ps1*z1+ps2*z2)
DH1Dps2<-la1*z2*exp(be1*d+ps1*z1+ps2*z2)

DH2Dla2<-exp(be2*d+et1*z1+et2*z2)
DH2Dbe2<-la2*d*exp(be2*d+et1*z1+et2*z2)
DH2Det1<-la2*z1*exp(be2*d+et1*z1+et2*z2)
DH2Det2<-la2*z2*exp(be2*d+et1*z1+et2*z2)

DSDla1<--exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*t
DSDbe1<--d*la1*exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*t
DSDps1<--la1*z1*exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*t
DSDps2<--la1*z2*exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*t

DSDla2<--exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*t
DSDbe2<--d*la2*exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*t
DSDet1<--la2*z1*exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*t
DSDet2<--la2*z2*exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*t

D2H1Dla1la1<-0
D2H1Dla1be1<-d*exp(be1*d+ps1*z1+ps2*z2)
D2H1Dla1ps1<-z1*exp(be1*d+ps1*z1+ps2*z2)
D2H1Dla1ps2<-z2*exp(be1*d+ps1*z1+ps2*z2)
D2H1Dbe1be1<-la1*d^2*exp(be1*d+ps1*z1+ps2*z2)
D2H1Dbe1ps1<-la1*z1*d*exp(be1*d+ps1*z1+ps2*z2)
D2H1Dbe1ps2<-la1*z2*d*exp(be1*d+ps1*z1+ps2*z2)
D2H1Dps1ps1<-la1*z1^2*exp(be1*d+ps1*z1+ps2*z2)
D2H1Dps1ps2<-la1*z1*z2*exp(be1*d+ps1*z1+ps2*z2)
D2H1Dps2ps2<-la1*z2^2*exp(be1*d+ps1*z1+ps2*z2)

D2H2Dla2la2<-0
D2H2Dla2be2<-d*exp(be2*d+et1*z1+et2*z2)
D2H2Dla2et1<-z1*exp(be2*d+et1*z1+et2*z2)
D2H2Dla2et2<-z2*exp(be2*d+et1*z1+et2*z2)
D2H2Dbe2be2<-la2*d^2*exp(be2*d+et1*z1+et2*z2)
D2H2Dbe2et1<-la2*z1*d*exp(be2*d+et1*z1+et2*z2)
D2H2Dbe2et2<-la2*z2*d*exp(be2*d+et1*z1+et2*z2)
D2H2Det1et1<-la2*z1^2*exp(be2*d+et1*z1+et2*z2)
D2H2Det1et2<-la2*z1*z2*exp(be2*d+et1*z1+et2*z2)
D2H2Det2et2<-la2*z2^2*exp(be2*d+et1*z1+et2*z2)

DH1Dla2<-0
DH1Dbe2<-0
DH1Det1<-0
DH1Det2<-0

DH2Dla1<-0
DH2Dbe1<-0
DH2Dps1<-0
DH2Dps2<-0

D2H2Dla1la2<-0
D2H2Dla1be2<-0
```

```
D2H2Dla1et1<-0
D2H2Dla1et2<-0
D2H2Dbe1la2<-0
D2H2Dbe1be2<-0
D2H2Dbe1et1<-0
D2H2Dbe1et2<-0
D2H2Dps1la2<-0
D2H2Dps1be2<-0
D2H2Dps1et1<-0
D2H2Dps1et2<-0
D2H2Dps2la2<-0
D2H2Dps2be2<-0
D2H2Dps2et1<-0
D2H2Dps2et2<-0

D2H1Dla2la1<-0
D2H1Dla2be1<-0
D2H1Dla2ps1<-0
D2H1Dla2ps2<-0
D2H1Dbe2la1<-0
D2H1Dbe2be1<-0
D2H1Dbe2ps1<-0
D2H1Dbe2ps2<-0
D2H1Det1la1<-0
D2H1Det1be1<-0
D2H1Det1ps1<-0
D2H1Det1ps2<-0
D2H1Det2la1<-0
D2H1Det2be1<-0
D2H1Det2ps1<-0
D2H1Det2ps2<-0


D2SDla1la1<-exp(2*be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+2*ps1*z1+2*ps2*z2)*t^2
D2SDla1be1<--exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*t*d*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDla1ps1<--exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*t*z1*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDla1ps2<--exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*t*z2*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDla1la2<-exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDla1be2<-d*la2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDla1et1<-la2*z1*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDla1et2<-la2*z2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDbe1be1<--d*exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*la1*t*d*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDbe1ps1<--d*exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*la1*t*z1*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDbe1ps2<--d*exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*la1*t*z2*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDbe1la2<-d*la1*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDbe1be2<-d^2*la1*la2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDbe1et1<-d*la1*la2*z1*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDbe1et2<-d*la1*la2*z2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDps1ps1<--exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*la1*t*z1*z1*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDps1ps2<--exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*la1*t*z1*z2*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDps1la2<-la1*z1*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDps1be2<-d*la1*la2*z1*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDps1et1<-la1*la2*z1^2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDps1et2<-la1*la2*z1*z2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDps2ps2<--exp(be1*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+ps1*z1+ps2*z2)*la1*t*z2*z2*(1-la1*exp(be1*d+ps1*z1+ps2*z2)*t)
D2SDps2la2<-la1*z2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDps2be2<-d*la1*la2*z2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDps2et1<-la1*la2*z1*z2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDps2et2<-la1*la2*z2^2*exp(be1*d+be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+ps1*z1+et2*z2+ps2*z2)*t^2
D2SDla2la2<-exp(2*be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-la2*exp(be2*d+et1*z1+et2*z2))*t+2*et1*z1+2*et2*z2)*t^2
D2SDla2be2<--exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*t*d*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)
D2SDla2et1<--exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*t*z1*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)
D2SDla2et2<--exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*t*z2*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)
D2SDbe2be2<--d*exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*la2*t*d*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)
D2SDbe2et1<--d*exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*la2*t*z1*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)
D2SDbe2et2<--d*exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*la2*t*z2*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)
D2SDet1et1<--exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*la2*t*z1*z1*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)
D2SDet1et2<--exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*la2*t*z1*z2*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)
D2SDet2et2<--exp(be2*d+(-la1*exp(be1*d+ps1*z1+ps2*z2)-
la2*exp(be2*d+et1*z1+et2*z2))*t+et1*z1+et2*z2)*la2*t*z2*z2*(1-la2*exp(be2*d+et1*z1+et2*z2)*t)


DlDph<-delta*(1-deltaTil)*J1*((1-St)/(1-ph*(1-St))-1/(1-ph))+
delta*(1-deltaTil)*J2*((1-St)/(1-ph*(1-St))-1/(1-ph))+((1-delta)*(1-deltaTil)*(1-St))/(1-ph*(1-St))
DlDxi1<-delta*deltaTil*((J1*Jtil1)/xi1-(J1*Jtil2)/(1-xi1)-
(H1*(J1*Jtil1+J2*Jtil1))/(H1*xi1+H2*(1-xi2))+(H1*(J2*Jtil2+J1*Jtil2))/(H1*(1-xi1)+H2*xi2))
DlDxi2<-delta*deltaTil*((J2*Jtil2)/xi2-(J2*Jtil1)/(1-xi2)+(H2*(J2*Jtil1+J1*Jtil1))/(H1*xi1+H2*(1-xi2))-
(H2*(J2*Jtil2+J1*Jtil2))/(H1*(1-xi1)+H2*xi2))
DlDH1<--((delta*(1-deltaTil)*(H1*J2-H2*J1))/(H1*(H1+H2)))+(delta*deltaTil*J1*(Jtil1+Jtil2))/H1-
(delta*deltaTil*(J1+J2)*Jtil2*(1-xi1))/(H1*(1-xi1)+H2*xi2)-
(delta*deltaTil*(J1+J2)*Jtil1*xi1)/(H1*xi1+H2*(1-xi2))
DlDH2<--((delta*(1-deltaTil)*(H2*J1-H1*J2))/(H2*(H1+H2)))+(delta*deltaTil*J2*(Jtil1+Jtil2))/H2-
(delta*deltaTil*(J1+J2)*Jtil1*(1-xi2))/(H1*xi1+H2*(1-xi2))-
(delta*deltaTil*(J1+J2)*Jtil2*xi2)/(H1*(1-xi1)+H2*xi2)
DlDS<--delta*(1-deltaTil)*(ph/(1-ph*(1-St))+1/(1-St))*(J1+J2)+(1-delta)*(1-deltaTil)*(1/St-ph/(1-ph*(1-St)))

D2lDphph<-delta*(1-deltaTil)*((1-St)^2/(1-ph*(1-St))^2-1/(1-ph)^2)*(J1+J2)+
((1-delta)*(1-deltaTil)*(1-St)^2)/(1-ph*(1-St))^2
D2lDphxi1<-0
D2lDphxi2<-0
D2lDphH1<-0
D2lDphH2<-0
D2lDphS<--delta*(1-deltaTil)*(J1+J2)*(1/(1-ph*(1-St))+(ph*(1-St))/(1-ph*(1-St))^2)-
((1-delta)*(1-deltaTil))/(1+ph*(-1+St))^2

D2lDxi1xi1<-delta*deltaTil*(-((J1*Jtil2)/(1-xi1)^2)-
(J1*Jtil1)/xi1^2+H1^2/(H1*xi1+H2*(1-xi2))^2*(J1*Jtil1+J2*Jtil1)+
H1^2/(H1*(1-xi1)+H2*xi2)^2*(J2*Jtil2+J1*Jtil2))
D2lDxi1xi2<--delta*deltaTil*((H1*H2)/(H1*xi1+H2*(1-xi2))^2*(J1*Jtil1+J2*Jtil1)+(H1*H2)/
(H1*(1-xi1)+H2*xi2)^2*(J1*Jtil2+J2*Jtil2))
D2lDxi1H1<-delta*deltaTil*H2*(J1+J2)*((Jtil2*xi2)/(H1-H1*xi1+H2*xi2)^2-(Jtil1*(1-xi2))/
(H1*xi1+H2*(1-xi2))^2)
D2lDxi1H2<-delta*deltaTil*H1*(J1+J2)*((Jtil1*(1-xi2))/(H1*xi1+H2*(1-xi2))^2-(Jtil2*xi2)/
(H1*(1-xi1)+H2*xi2)^2)

D2lDxi1S<-0

D2lDxi2xi2<-delta*deltaTil*(H2^2/(H1*xi1+H2*(1-xi2))^2*(J2*Jtil1+J1*Jtil1)+
H2^2/(H1*(1-xi1)+H2*xi2)^2*(J1*Jtil2+J2*Jtil2)-(J2*Jtil2)/xi2^2-(J2*Jtil1)/(1-xi2)^2)
D2lDxi2H1<-delta*deltaTil*H2*(J1+J2)*((Jtil2*(1-xi1))/(H1*(-1+xi1)-H2*xi2)^2-(Jtil1*xi1)/(H2+H1*xi1-H2*xi2)^2)
D2lDxi2H2<-delta*deltaTil*H1*(J1+J2)*((Jtil1*xi1)/(H1*xi1+H2*(1-xi2))^2-(Jtil2*(1-xi1))/(H1*(1-xi1)+H2*xi2)^2)
D2lDxi2S<-0
```

```
D2lDH1H1<-(delta*(1-deltaTil)*(-2*H1*H2*J1-H2^2*J1+H1^2*J2))/(H1^2*(H1+H2)^2)-
(delta*deltaTil*J1*(Jtil1+Jtil2))/H1^2+(delta*deltaTil*(J1+J2)*Jtil1*xi1^2)/(H1*xi1+H2*(1-xi2))^2+
(delta*deltaTil*(J1+J2)*Jtil2*(1-xi1)^2)/(H1*(1-xi1)+H2*xi2)^2
D2lDH1H2<-(delta*(1-deltaTil)*(J1+J2))/(H1+H2)^2+
delta*deltaTil*(J1+J2)*((Jtil2*(1-xi1)*xi2)/(H1*(1-xi1)+H2*xi2)^2+
(Jtil1*xi1*(1-xi2))/(H1*xi1+H2*(1-xi2))^2)
D2lDH1S<-0

D2lDH2H2<-(delta*(1-deltaTil)*(H2^2*J1-H1*J2*(H1+2*H2)))/(H2^2*(H1+H2)^2)-
(delta*deltaTil*J2*(Jtil1+Jtil2))/H2^2+delta*deltaTil*(J1+J2)*((Jtil2*xi2^2)/
(H1*(1-xi1)+H2*xi2)^2+(Jtil1*(-1+xi2)^2)/(H1*xi1+H2*(1-xi2))^2)
D2lDH2S<-0

D2lDSS<-delta*(1-deltaTil)*(ph^2/(1-ph*(1-St))^2-1/(1-St)^2)*(J1+J2)+
(1-delta)*(1-deltaTil)*(ph^2/(1-ph*(1-St))^2-1/St^2)

D2lDH1ph<-0
D2lDH1xi1<-D2lDxi1H1
D2lDH1xi2<-D2lDxi2H1
D2lDH2ph<-0
D2lDH2xi1<-D2lDxi1H2
D2lDH2xi2<-D2lDxi2H2
D2lDSph<-D2lDphS
D2lDSxi1<-0
D2lDSxi2<-0


I11<--sum(DlDH1*D2H1Dla1la1+DlDS*D2SDla1la1+D2lDH1H1*(DH1Dla1)^2+D2lDSS*(DSDla1)^2+2*D2lDH1S*DH1Dla1*DSDla1)
I22<--sum(DlDH1*D2H1Dbe1be1+DlDS*D2SDbe1be1+D2lDH1H1*(DH1Dbe1)^2+D2lDSS*(DSDbe1)^2+2*D2lDH1S*DH1Dbe1*DSDbe1)
I33<--sum(DlDH1*D2H1Dps1ps1+DlDS*D2SDps1ps1+D2lDH1H1*(DH1Dps1)^2+D2lDSS*(DSDps1)^2+2*D2lDH1S*DH1Dps1*DSDps1)
I44<--sum(DlDH1*D2H1Dps2ps2+DlDS*D2SDps2ps2+D2lDH1H1*(DH1Dps2)^2+D2lDSS*(DSDps2)^2+2*D2lDH1S*DH1Dps2*DSDps2)
I55<--sum(DlDH2*D2H2Dla2la2+DlDS*D2SDla2la2+D2lDH2H2*(DH2Dla2)^2+D2lDSS*(DSDla2)^2+2*D2lDH2S*DH2Dla2*DSDla2)
I66<--sum(DlDH2*D2H2Dbe2be2+DlDS*D2SDbe2be2+D2lDH2H2*(DH2Dbe2)^2+D2lDSS*(DSDbe2)^2+2*D2lDH2S*DH2Dbe2*DSDbe2)
I77<--sum(DlDH2*D2H2Det1et1+DlDS*D2SDet1et1+D2lDH2H2*(DH2Det1)^2+D2lDSS*(DSDet1)^2+2*D2lDH2S*DH2Det1*DSDet1)
I88<--sum(DlDH2*D2H2Det2et2+DlDS*D2SDet2et2+D2lDH2H2*(DH2Det2)^2+D2lDSS*(DSDet2)^2+2*D2lDH2S*DH2Det2*DSDet2)
I99<--sum(D2lDphph)
I1010<--sum(D2lDxi1xi1)
I1111<--sum(D2lDxi2xi2)

I12<--sum(DlDH1*D2H1Dla1be1+DH1Dla1*(D2lDH1H1*DH1Dbe1+D2lDH1S*DSDbe1)+
DSDla1*(D2lDSS*DSDbe1+D2lDH1S*DH1Dbe1)+D2SDla1be1*DlDS)
I13<--sum(DlDH1*D2H1Dla1ps1+DH1Dla1*(D2lDH1H1*DH1Dps1+D2lDH1S*DSDps1)+
DSDla1*(D2lDSS*DSDps1+D2lDH1S*DH1Dps1)+D2SDla1ps1*DlDS)
I14<--sum(DlDH1*D2H1Dla1ps2+DH1Dla1*(D2lDH1H1*DH1Dps2+D2lDH1S*DSDps2)+
DSDla1*(D2lDSS*DSDps2+D2lDH1S*DH1Dps2)+D2SDla1ps2*DlDS)
I15<--sum(DlDH2*D2H2Dla1la2+DH2Dla2*(D2lDH2H2*DH2Dla1+D2lDH2S*DSDla1+
D2lDH1H2*DH1Dla1)+DSDla2*(D2lDSS*DSDla1+D2lDH2S*DH2Dla1+D2lDH1S*DH1Dla1)+D2SDla1la2*DlDS)
I16<--sum(DlDH2*D2H2Dla1be2+DH2Dbe2*(D2lDH2H2*DH2Dla1+D2lDH2S*DSDla1+
D2lDH1H2*DH1Dla1)+DSDbe2*(D2lDSS*DSDla1+D2lDH2S*DH2Dla1+D2lDH1S*DH1Dla1)+D2SDla1be2*DlDS)
I17<--sum(DlDH2*D2H2Dla1et1+DH2Det1*(D2lDH2H2*DH2Dla1+D2lDH2S*DSDla1+
D2lDH1H2*DH1Dla1)+DSDet1*(D2lDSS*DSDla1+D2lDH2S*DH2Dla1+D2lDH1S*DH1Dla1)+D2SDla1et1*DlDS)
I18<--sum(DlDH2*D2H2Dla1et2+DH2Det2*(D2lDH2H2*DH2Dla1+D2lDH2S*DSDla1+
D2lDH1H2*DH1Dla1)+DSDet2*(D2lDSS*DSDla1+D2lDH2S*DH2Dla1+D2lDH1S*DH1Dla1)+D2SDla1et2*DlDS)
I19<--sum(D2lDH1ph*DH1Dla1+D2lDSph*DSDla1)
I110<--sum(D2lDH1xi1*DH1Dla1+D2lDSxi1*DSDla1)
I111<--sum(D2lDH1xi2*DH1Dla1+D2lDSxi2*DSDla1)

I23<--sum(DlDH1*D2H1Dbe1ps1+DH1Dbe1*(D2lDH1H1*DH1Dps1+D2lDH1S*DSDps1)+
DSDbe1*(D2lDSS*DSDps1+D2lDH1S*DH1Dps1)+D2SDbe1ps1*DlDS)
I24<--sum(DlDH1*D2H1Dbe1ps2+DH1Dbe1*(D2lDH1H1*DH1Dps2+D2lDH1S*DSDps2)+
DSDbe1*(D2lDSS*DSDps2+D2lDH1S*DH1Dps2)+D2SDbe1ps2*DlDS)
I25<--sum(DlDH2*D2H2Dbe1la2+DH2Dla2*(D2lDH2H2*DH2Dbe1+D2lDH2S*DSDbe1+
D2lDH1H2*DH1Dbe1)+DSDla2*(D2lDSS*DSDbe1+D2lDH2S*DH2Dbe1+D2lDH1S*DH1Dbe1)+D2SDbe1la2*DlDS)
I26<--sum(DlDH2*D2H2Dbe1be2+DH2Dbe2*(D2lDH2H2*DH2Dbe1+D2lDH2S*DSDbe1+
D2lDH1H2*DH1Dbe1)+DSDbe2*(D2lDSS*DSDbe1+D2lDH2S*DH2Dbe1+D2lDH1S*DH1Dbe1)+D2SDbe1be2*DlDS)
I27<--sum(DlDH2*D2H2Dbe1et1+DH2Det1*(D2lDH2H2*DH2Dbe1+D2lDH2S*DSDbe1+
D2lDH1H2*DH1Dbe1)+DSDet1*(D2lDSS*DSDbe1+D2lDH2S*DH2Dbe1+D2lDH1S*DH1Dbe1)+D2SDbe1et1*DlDS)
I28<--sum(DlDH2*D2H2Dbe1et2+DH2Det2*(D2lDH2H2*DH2Dbe1+D2lDH2S*DSDbe1+
D2lDH1H2*DH1Dbe1)+DSDet2*(D2lDSS*DSDbe1+D2lDH2S*DH2Dbe1+D2lDH1S*DH1Dbe1)+D2SDbe1et2*DlDS)
I29<--sum(D2lDH1ph*DH1Dbe1+D2lDSph*DSDbe1)
I210<--sum(D2lDH1xi1*DH1Dbe1+D2lDSxi1*DSDbe1)
I211<--sum(D2lDH1xi2*DH1Dbe1+D2lDSxi2*DSDbe1)
```

```
I34<--sum(DlDH1*D2H1Dps1ps2+DH1Dps1*(D2lDH1H1*DH1Dps2+D2lDH1S*DSDps2)+
DSDps1*(D2lDSS*DSDps2+D2lDH1S*DH1Dps2)+D2SDps1ps2*DlDS)
I35<--sum(DlDH2*D2H2Dps1la2+DH2Dla2*(D2lDH2H2*DH2Dps1+D2lDH2S*DSDps1+D2lDH1H2*DH1Dps1)+
DSDla2*(D2lDSS*DSDps1+D2lDH2S*DH2Dps1+D2lDH1S*DH1Dps1)+D2SDps1la2*DlDS)
I36<--sum(DlDH2*D2H2Dps1be2+DH2Dbe2*(D2lDH2H2*DH2Dps1+D2lDH2S*DSDps1+D2lDH1H2*DH1Dps1)+
DSDbe2*(D2lDSS*DSDps1+D2lDH2S*DH2Dps1+D2lDH1S*DH1Dps1)+D2SDps1be2*DlDS)
I37<--sum(DlDH2*D2H2Dps1et1+DH2Det1*(D2lDH2H2*DH2Dps1+D2lDH2S*DSDps1+D2lDH1H2*DH1Dps1)+
DSDet1*(D2lDSS*DSDps1+D2lDH2S*DH2Dps1+D2lDH1S*DH1Dps1)+D2SDps1et1*DlDS)
I38<--sum(DlDH2*D2H2Dps1et2+DH2Det2*(D2lDH2H2*DH2Dps1+D2lDH2S*DSDps1+D2lDH1H2*DH1Dps1)+
DSDet2*(D2lDSS*DSDps1+D2lDH2S*DH2Dps1+D2lDH1S*DH1Dps1)+D2SDps1et2*DlDS)
I39<--sum(D2lDH1ph*DH1Dps1+D2lDSph*DSDps1)
I310<--sum(D2lDH1xi1*DH1Dps1+D2lDSxi1*DSDps1)
I311<--sum(D2lDH1xi2*DH1Dps1+D2lDSxi2*DSDps1)

I45<--sum(DlDH2*D2H2Dps2la2+DH2Dla2*(D2lDH2H2*DH2Dps2+D2lDH2S*DSDps2+D2lDH1H2*DH1Dps2)+
DSDla2*(D2lDSS*DSDps2+D2lDH2S*DH2Dps2+D2lDH1S*DH1Dps2)+D2SDps2la2*DlDS)
I46<--sum(DlDH2*D2H2Dps2be2+DH2Dbe2*(D2lDH2H2*DH2Dps2+D2lDH2S*DSDps2+D2lDH1H2*DH1Dps2)+
DSDbe2*(D2lDSS*DSDps2+D2lDH2S*DH2Dps2+D2lDH1S*DH1Dps2)+D2SDps2be2*DlDS)
I47<--sum(DlDH2*D2H2Dps2et1+DH2Det1*(D2lDH2H2*DH2Dps2+D2lDH2S*DSDps2+D2lDH1H2*DH1Dps2)+
DSDet1*(D2lDSS*DSDps2+D2lDH2S*DH2Dps2+D2lDH1S*DH1Dps2)+D2SDps2et1*DlDS)
I48<--sum(DlDH2*D2H2Dps2et2+DH2Det2*(D2lDH2H2*DH2Dps2+D2lDH2S*DSDps2+D2lDH1H2*DH1Dps2)+
DSDet2*(D2lDSS*DSDps2+D2lDH2S*DH2Dps2+D2lDH1S*DH1Dps2)+D2SDps2et2*DlDS)
I49<--sum(D2lDH1ph*DH1Dps2+D2lDSph*DSDps2)
I410<--sum(D2lDH1xi1*DH1Dps2+D2lDSxi1*DSDps2)
I411<--sum(D2lDH1xi2*DH1Dps2+D2lDSxi2*DSDps2)

I56<--sum(DlDH2*D2H2Dla2be2+DH2Dla2*(D2lDH2H2*DH2Dbe2+D2lDH2S*DSDbe2)+
DSDla2*(D2lDSS*DSDbe2+D2lDH2S*DH2Dbe2)+D2SDla2be2*DlDS)
I57<--sum(DlDH2*D2H2Dla2et1+DH2Dla2*(D2lDH2H2*DH2Det1+D2lDH2S*DSDet1)+
DSDla2*(D2lDSS*DSDet1+D2lDH2S*DH2Det1)+D2SDla2et1*DlDS)
I58<--sum(DlDH2*D2H2Dla2et2+DH2Dla2*(D2lDH2H2*DH2Det2+D2lDH2S*DSDet2)+
DSDla2*(D2lDSS*DSDet2+D2lDH2S*DH2Det2)+D2SDla2et2*DlDS)
I59<--sum(D2lDH2ph*DH2Dla2+D2lDSph*DSDla2)
I510<--sum(D2lDH2xi1*DH2Dla2+D2lDSxi1*DSDla2)
I511<--sum(D2lDH2xi2*DH2Dla2+D2lDSxi2*DSDla2)

I67<--sum(DlDH2*D2H2Dbe2et1+DH2Dbe2*(D2lDH2H2*DH2Det1+D2lDH2S*DSDet1)+
DSDbe2*(D2lDSS*DSDet1+D2lDH2S*DH2Det1)+D2SDbe2et1*DlDS)
I68<--sum(DlDH2*D2H2Dbe2et2+DH2Dbe2*(D2lDH2H2*DH2Det2+D2lDH2S*DSDet2)+
DSDbe2*(D2lDSS*DSDet2+D2lDH2S*DH2Det2)+D2SDbe2et2*DlDS)
I69<--sum(D2lDH2ph*DH2Dbe2+D2lDSph*DSDbe2)
I610<--sum(D2lDH2xi1*DH2Dbe2+D2lDSxi1*DSDbe2)
I611<--sum(D2lDH2xi2*DH2Dbe2+D2lDSxi2*DSDbe2)

I78<--sum(DlDH2*D2H2Det1et2+DH2Det1*(D2lDH2H2*DH2Det2+D2lDH2S*DSDet2)+
DSDet1*(D2lDSS*DSDet2+D2lDH2S*DH2Det2)+D2SDet1et2*DlDS)
I79<--sum(D2lDH2ph*DH2Det1+D2lDSph*DSDet1)
I710<--sum(D2lDH2xi1*DH2Det1+D2lDSxi1*DSDet1)
I711<--sum(D2lDH2xi2*DH2Det1+D2lDSxi2*DSDet1)

I89<--sum(D2lDH2ph*DH2Det2+D2lDSph*DSDet2)
I810<--sum(D2lDH2xi1*DH2Det2+D2lDSxi1*DSDet2)
I811<--sum(D2lDH2xi2*DH2Det2+D2lDSxi2*DSDet2)

I910<--sum(D2lDphxi1)
I911<--sum(D2lDphxi2)

I1011<--sum(D2lDxi1xi2)

I<-rbind(c(I11,I12,I13,I14,I15,I16,I17,I18,I19,I110,I111),
         c(I12,I22,I23,I24,I25,I26,I27,I28,I29,I210,I211),
         c(I13,I23,I33,I34,I35,I36,I37,I38,I39,I310,I311),
         c(I14,I24,I34,I44,I45,I46,I47,I48,I49,I410,I411),
         c(I15,I25,I35,I45,I55,I56,I57,I58,I59,I510,I511),
         c(I16,I26,I36,I46,I56,I66,I67,I68,I69,I610,I611),
         c(I17,I27,I37,I47,I57,I67,I77,I78,I79,I710,I711),
         c(I18,I28,I38,I48,I58,I68,I78,I88,I89,I810,I811),
         c(I19,I29,I39,I49,I59,I69,I79,I89,I99,I910,I911),
         c(I110,I210,I310,I410,I510,I610,I710,I810,I910,I1010,I1011),
         c(I111,I211,I311,I411,I511,I611,I711,I811,I911,I1011,I1111))

return(I)
}
```

Main Simulation Code

```
library(survival)
numsim<-1000

Lambda1<-0.05
Beta1<-0
Psi1<-0.4
Psi2<-0
Lambda2<-0.06
Beta2<-0
Eta1<-0.2
Eta2<-0.5
Phi<-0.99
Xi1<-0.9
Xi2<-0.95
Mu<-.0055
Pd<-0.4
Pz1<-0.6
n<-10000
m<-1000
pM1<-0.5

parList<-c(Lambda1,Beta1, Psi1, Psi2, Lambda2, Beta2,
Xi1, Xi2, Mu, Pd, Pz1, n, m, pM1)
source("C:\\Users\\CG\\Dropbox\\MISCLASS OR REWRITE\\Failure time\\
COMP RISKS\\competing risks MC delta and xi Information Matrices 9-27-2014.r")

res<-{}
resC<-{}
Vars<-{}
resPerfClass<-{}
VarsPerfClass<-{}

SeedS<-runif(numsim,0,10000000)

for(i in 1:numsim) {
P<-as.numeric(PARS[k,])

set.seed(SeedS[i])
dat<-datagenCompRisks.8par(parList)

d<-dat$D
val<-dat$v

d<-as.data.frame(dat$D)
v<-as.data.frame(dat$v)

mle<-nlminb(c(1,0,0,0,1,0,0,0,0.5,0.5,0.5),LL8,
lower=c(0.000000001,-Inf,-Inf,-Inf,0.0000000001,-Inf,-Inf,-Inf,0.0000000001,0.0000000001,0.0000000001),
upper=c(Inf,Inf,Inf,Inf,Inf,Inf,Inf,Inf,0.9999999999,0.9999999999,0.9999999999),dat=d,val=val)$par
Vars<-rbind(Vars,diag(tryCatch(solve(IOtype(d,mle)+IVtype(val,mle)),error=function(e) e=matrix(10000,11,11))))
res<-rbind(res,mle)

PerfClassReg1<-survreg(Surv(d[,1],d[,6]==1) ~ d[,8]+d[,9]+d[,10],dist="exponential")
PerfClassReg2<-survreg(Surv(d[,1],d[,6]==2) ~ d[,8]+d[,9]+d[,10],dist="exponential")

resPerfClass<-rbind(resPerfClass,c(1/exp(PerfClassReg1$coefficients[1]),
-PerfClassReg1$coefficients[2:4],1/exp(PerfClassReg2$coefficients[1]),-PerfClassReg2$coefficients[2:4]))
VarsPerfClass<-rbind(VarsPerfClass,c(PerfClassReg1$var[1]/exp(2*PerfClassReg1$coefficients[1]),
diag(PerfClassReg1$var)[2:4],PerfClassReg2$var[1]/exp(2*PerfClassReg2$coefficients[1]),
diag(PerfClassReg2$var)[2:4]))
}

True<-parList
CI<-{}
for(l in 1:11) {
numberNA<-which(suppressWarnings(is.na(sqrt(Vars[,l]))==FALSE))
CI<-c(CI,length(which(res[numberNA,l]-1.96*sqrt(Vars[numberNA,l])<
rep(True[l],length(numberNA))&
res[numberNA,l]+1.96*sqrt(Vars[numberNA,l]) >
rep(True[l],length(numberNA))
))/length(numberNA))
}
CItil<-{}
for(l in 1:8)
CItil<-c(CItil,length(which(resPerfClass[,l]-1.96*sqrt(VarsPerfClass[,l])<rep(True[l+1],numsim)&
resPerfClass[,l]+1.96*sqrt(VarsPerfClass[,l]) >rep(True[l+1],numsim)))/numsim)
CItil<-CItil
gammaTil<-apply(resPerfClass,2,mean)
gammaTilSE<-sqrt(apply(VarsPerfClass,2,mean))
gammaHat<-apply(res,2,mean)[1:11]
gammaHatSE<-sqrt(apply(Vars[which(Vars[,1]!=10000),],2,mean))[1:11]
MCSE<-sqrt(apply(res,2,var))
Results<-as.data.frame(cbind(as.character(True[1:11]),c(as.character(round(gammaTil,5)),"-","-","-"),
c(as.character(round(gammaTilSE,5)),"-","-","-"),c(as.character(CItil),"-","-","-"),
round(gammaHat,5),round(gammaHatSE,5),round(MCSE,5),round(CI,5)))
colnames(Results)<-c("True","gamTil","gamTilSE","CItil","gamHat","gamHatSE","MCSE","CI")
rownames(Results)<-c("Lambda1","Beta1","Psi1","Psi2","Lambda2","Beta2","Eta1","Eta2","Phi","Xi1","Xi2")
```

Appendix D

Additional Tables

A large number of simulation results were generated for the work presented in this thesis. However, to limit its length we will provide a second accompanying document with these additional tables upon request from the author (christophergravel@cmail.carleton.ca).

Bibliography

[1] Agresti, A. (2002). Categorical Data Analysis. Second Edition, New York: Wiley.

[2] Barron, B. A. (1977). The effects of misclassification on the estimation of relative risk. *Biometrics* 33, 414-418.

[3] Beyersmann J, Latouche A, Buchholz A, Schumacher M. (2009). Simulating competing risks data in survival analysis. *Stat Med.* 28(6):956-71.

[4] Carroll RJ, Ruppert D, Stefanski LA (2006). Measurement Error in Nonlinear Models. Second Edition, Chapman and Hall: London.

[5] Copeland, K. T., Checkoway, H., McMichael, A. J., and Holbrook, R. H. (1977). Bias due to misclassification in the estimation of relative risk. *American Journal of Epidemiology.* 105, 488-495.

[6] Drahos J., Vanwormer J. J., Greenlee R. T., Landgren O., Koshiol J. (2013). Accuracy of ICD-9-CM codes in identifying infections of pneumonia and herpes simplex virus in administrative data. *Annals of Epidemiology.* 23: 291-293.

[7] Espeland, M. A. and Hui, S. L. (1987). A general approach to analyzing epidemiologic data that contain misclassification errors. *Biometrics.* 43, 1001-1012.

[8] Ghosh, P. and Dewanji, A. (2011). Analysis of spontaneous adverse drug reaction (ADR) reports using supplementary information. *Statistics in Medicine*. 30, 2040-2055.

[9] Goldberg, J. D. (1975). The Effects of Misclassification on the Bias in the Difference between Two Proportions and the Relative Odds in the Fourfold Table. *Journal of the American Statistical Association*. 70, 561-567.

[10] Greenland, S. (1982). The effect of misclassification in matched-pair case-control studies. *American Journal of Epidemiology*. 116, 402-406.

[11] Greenland, S. and Kleinbaum, D.G. (1983). Correcting for misclassification in two-way tables and matched-pair studies. *International Journal of Epidemiology*. 12, 93-97.

[12] Greenland, S. (1988). Variance Estimation for Epidemiologic Effect Estimates Under Misclassification. *Statistics in Medicine*. 7, 745-757.

[13] Greenland, S (2008). Maximum-likelihood and closed-form estimators of Epidemiologic Measures Under Misclassification. *Journal of Statistical Planning and Inference*. 138, 528-538.

[14] B.A., Stamey J.D., Young D.M., Ryden D.J. (2009). An Alterntive Derivation of the Multi-Parameter Cramer-Rao Inequality. *Math. Scientist*. 34, 20-24

[15] Harville, D.A. (1997). Matrix Algebra From a Statistician's Perspective. New York: Springer-Verlag.

[16] Häyrinen K., Saranto K., Nykänen P. (2008). Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *Int J Med Inform* 77(5):291-304.

[17] Hirose Y. (2005). Efficiency of the semi-parametric maximum likelihood estimator in generalized case–control studies. Ph.D. Thesis, University of Auckland, 67-94.

[18] A. Kagan, P.J. Smith (2001). Multivariate normal distributions, Fisher information and matrix inequalities. *International Journal of Mathematical Education in Science and Technology.* 32:1, 91-96.

[19] Kalbfleisch, J.D. and Prentice, R.L. (2002). The Statistical Analysis of Failure Time Data, 2nd edition. John Wiley and Sons, New York.

[20] Korn, E. L. (1981). Hierarchical log-linearmodels not preserved by classification error. *Journal of the American Statistical Association.* 76, 110-113.

[21] Lehmann EL, Casella G (1998). Theory of Point Estimation (2nd edn). Springer: Berlin.

[22] Levine P. J., Elman M. R., Kullar R., Townes J. M., Bearden D. T., Vilches-Tran R., McClellan I., McGregor J. C. (2013). Use of electronic health record data to identify skin and soft tissue infections in primary care settings: a validation study. *BMC Infectious Diseases.* 13: 171

[23] Lyles, R. H. (2002). A Note on Estimating Crude Odds Ratios in Case-Control Studies with Differentially Misclassified Exposure. *Biometrics.* 58:1034-1037.

[24] Lyles, R. H., Allen, A. S. (2003). Missing data in the 2 × 2 table: patterns and likelihood-based analysis for cross-sectional studies with supplemental sampling. *Statistics in Medicine.* 22:517-534.

[25] Lyles RH, Tang L, Superak HM, King CC, Celentano DD, Lo Y, Sobel JD (2011). Validation data-based adjustments for outcome misclassification in logistic regression. *Epidemiology*; 22:589–98.

[26] A. Magaret (2008). Incorporating validation subsets into discrete proportional hazards models for mismeasured outcomes. *Statist. Med.* 27:5456–5470.

[27] Magder LS, Hughes JP (1997). Logistic regression when the outcome is measured with uncertainty. *Am J Epidemiol.* 146:195–203.

[28] Marshall, R. J. (1990). Validation study methods for estimating proportions and odds ratios with misclassified data. *Journal of Clinical Epidemiology.* 43, 941-947.

[29] K. McKeown, N.P. Jewell (2010). Misclassification of current status data. *Lifetime Data Anal.* 16:215–230.

[30] Meier AS, Richardson BA, Hughes JP (2003). Discrete proportional hazards models for mismeasured outcomes. *Biometrics.* 59(4):947–954.

[31] Montgomery D.C. (2005). Design and Analysis of Experiments. Sixth Edition, New Jersey: Wiley.

[32] Morrissey MJ, Spiegelman D (1999). Matrix methods for estimating odds-ratios with misclassified exposure data: extensions and comparisons. *Biometrics.* 55:338–344.

[33] Nicholson A., Tate R. T., Koeling R. and Cassell J. A. (2011). What does validation of cases in electronic record databases mean? The potential contribution of free text. *Pharmacoepidemiology and Drug Safety.* 20: 321-324.

[34] Richardson, B.A., Hughes, J.P. (2000). Product limit estimation for infectious disease data when the diagnostic test for the outcome is measured with uncertainty. *Biostatistics.* 1,3:341-354.

[35] Rogan, W. J. and Gladen, B. (1978). Estimating the prevalence from the results of a screening test. *American Journal of Epidemiology.* 107, 71-76.

[36] Van Rompaye B, Jaffar S, and Goetghebeur E (2010). Design and testing for clinical trials faced with misclassified causes of death. *Biostatistics.* 11:546–558.

[37] Roux, E., Thiessard, F., Fourrier, A., Begaud, B., and Tubert-Bitter, P (2005). Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance. *IEEE Trans.Inf.Technol.Biomed.* 9: 518-527.

[38] Tenenbein, A. (1970). A Double Sampling Scheme for Estimating from Misclassified Binomial Data. *Journal of the American Statistical Association.* 65, 1350-1361.

[39] Whaley, F. S., Quade, D., and Haley, R. W. (1980). Effects of method error on the power of a statistical test. *American Journal of Epidemiology.* 111, 534-542.

[40] http://www.who.int/classifications/icd/en/ Accessed: 04-09-2013.