Modeling and Prediction of Disease Processes Subject to Intermittent Observation

by

Ying Wu

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

Waterloo, Ontario, Canada, 2016

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

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

Abstract

This thesis is concerned with statistical modeling and prediction of disease processes subject to intermittent observation. Times of disease progression are interval-censored when progression status is only known at a series of assessment times. This situation arises routinely in clinical trials and cohort studies when events of interest are only detectable upon imaging, based on blood tests, or upon careful clinical examination. The work that follows is motivated by the study of demographic, genetic and clinical data available from the University of Toronto Psoriasis Registry and the University of Toronto Psoriatic Arthritis Registry, each involving cohorts of several hundred patients with the respective diseases.

Chapter 2 deals with the problem of selecting important prognostic biomarkers from a large set of candidates biomarkers when the status with respect to an event of interest (e.g. disease progression) is only known at irregularly spaced and individual-specific assessment times. Penalized regression techniques (e.g. LASSO, adaptive LASSO and SCAD) are adapted to deal with the interval-censored event times arising from this observation scheme. An expectation-maximization algorithm is developed which is demonstrated to perform well in extensive simulation studies involving independent and correlated continuous and binary covariates. Application to the motivating study of the development of arthritis mutilans in patients with psoriatic arthritis is given and several important human leukocyte antigen (HLA) variables are identified for further investigation. Extensions of this algorithm are developed for settings in which data from different sources with distinct disease-related entry conditions are to be synthesized. The extended Turnbull-type expectation-maximization algorithm is based on a complete data likelihood which incorporates missing information from individuals not meeting the entry criteria of the respective registries. Simulation studies demonstrate good empirical performance and an application to the motivating study identifies HLA markers associated with the onset of psoriatic

arthritis among individuals with psoriasis. This analysis is carried out using data from a psoriasis registry in which the times to psoriatic arthritis are left-truncated, and psoriatic arthritis registry in which the onset times are right-truncated.

Chapter 3 deals with the challenge of assessing the accuracy of a predictive model when response times are interval-censored. Inverse probability weighted (IPW) and augmented inverse probability weighted (AIPW) estimators of predictive accuracy are developed and evaluated based on the mean prediction error and the area under the receiver operating characteristic curve. The weights are estimated from a multistate model which jointly considers the event process, the inspection process, and the right-censoring processes. We investigate the performance of the proposed methods by simulation and illustrate their application in the context of a motivating rheumatology study in which HLA markers are used for predicting disease progression in psoriatic arthritis.

A two-phase model is developed in Chapter 4 for chronic diseases which feature an indolent phase followed by a phase with more active disease resulting in progression and damage. The time-scales for the intensity functions for the active phase are more naturally based on the time since the start of the active phase, corresponding to a semi-Markov formulation. In cohort studies for which the disease status is only known at a series of clinical assessment times, transition times are interval-censored which means the time origin for phase II is interval-censored. Weakly parametric models with piecewise constant baseline hazard and rate functions are specified and an expectation-maximization algorithm is described for model fitting. A computationally faster two-stage estimation procedure is also developed and the asymptotic variances of the resulting estimators are derived. Simulation studies examining the performance of the proposed model show good performance under both maximum likelihood and two-stage estimation. An application to data from the motivating study of disease progression in psoriatic arthritis illustrates the procedure,

and identifies new human leukocyte antigens associated with the duration of the indolent phase, and others associated with disease progression in the active phase.

Open problems and topics for ongoing and future research are discussed in Chapter 5.

Acknowledgements

Foremost, I would like to express my deepest gratitude to my supervisor Dr. Richard Cook for the continuous support and encouragement of my study and research, for his guidance, patience, valuable insight and advice, immense knowledge and experience. I could not have come this far or even started without his motivation, encouragement and trust in my abilities. I am so grateful to him for bringing great opportunities for me and providing financial support to facilitate my research. I am very lucky and grateful to work with him for the past few years as a student and a research assistant.

I would also like to express my sincere gratitude to my thesis committee. I would like to thank Dr. Jerry Lawless and Dr. Leilei Zeng for their insightful comments and helpful suggestions from the beginning of thesis proposal to the end of defence. I would also like to thank Dr. Thierry Duchesne from University of Laval and Dr. Hossein Abouee-Mehrizi from Department of Management Sciences for taking their time to serve as the external and internal-external member of my thesis committee, for their expertise from different aspects and inspiring questions to perfect my thesis. A very special thanks goes to Dr. Leilei Zeng, for sharing experience in both research and career building, for creating opportunities for collaboration to widen my knowledge.

I must also acknowledge Dr. Dafna Gladman, Dr. Lihi Eder, Dr. Vinod Chandrad and the team from the University of Toronto Psoriatic Arthritis Clinic for the use of the datasets. A very special thanks goes to Ker-Ai Lee, for her encouragement and help in every aspect of my research. Appreciation also goes to Shared Hierarchical Academic Research Computing Network (SHARCNET) and Compute Canada for facilitating the computational work in my thesis.

I wish to acknowledge the faculty, students and staff of the Department of Statistics

and Actuarial Science. In particular I wish to thank Mary Lou Dufton and Marg Feeney for their support and help, as well as all the students in the department for sharing this journey together.

Last but not the least, I would like to thank my friends and family, who are always standing by me and encouraging me to be myself and pursue my passion and dream.

Dedication

This is dedicated to my parents and grandparents.

Table of Contents

List of Tables **xiv**

List of Figures **xix**

1 Introduction **1**

1.1 Motivating Research Program 1

1.1.1 Overview 1

1.1.2 Psoriasis and Psoriatic Arthritis 2

1.1.3 Arthritis Mutilans 5

1.2 General Introduction to Research Topics 5

1.2.1 Penalized Regression for Interval-Censored Times of Disease Progression 6

1.2.2 Assessing the Accuracy of Predictive Models with Interval-Censored Data 7

1.2.3 A Two-Phase Model for Chronic Disease Processes Under Intermittent Inspection 7

2 Penalized Regression for Interval-Censored Times of Disease Progression 9

2.1 Introduction 9

2.1.1 Variable Selection and Penalized Regression 9

2.1.2 Prognostic HLA Markers in Psoriatic Arthritis 12

2.2 Variable Selection with Interval-Censored Data 15

2.2.1 Notation and the Penalized Complete Data Likelihood 15

2.2.2 An Expectation-Maximization Algorithm 17

2.3 Design and Interpretation of Simulation Studies 21

2.4 HLA Markers and Risk of Arthritis Mutilans 29

2.5 Summary of Findings on Penalized Regression for Interval-Censored Data . 33

2.6 Penalized Regression for Truncated and Censored Data 35

2.6.1 The Motivating Study and Sample Selection Conditions 35

2.6.2 A Turnbull-Type EM Algorithm 38

2.6.3 Simulation Studies and Application to the Psoriasis and Psoriatic Arthritis Registries 41

2.6.4 Discussion 45

Appendix 2.A Supplementary Simulation Studies 46

Appendix 2.B Comparison of Methods for Choosing the Optimal Tuning Parameter 51

Appendix 2.C Variance Estimation 55

3 Assessing the Accuracy of Predictive Models with Interval-Censored Data **58**

3.1 Introduction 58

3.1.1 Overview 58

3.1.2 Estimating Prediction Error 60

3.1.3 Estimating Prediction Error for Censored Data 63

3.2 Prediction for Interval-Censored Data 65

3.2.1 Notation and Formulation of Observation Process Models 66

3.2.2 Inverse Probability Weighted Estimator 68

3.2.3 Augmented Inverse Probability Weighted Estimator 70

3.2.4 ROC Curves and the Area Under the Curve 72

3.3 Simulation Studies 73

3.3.1 Design and Results of Studies for Poisson Processes 73

3.3.2 Design and Results of Studies for Renewal Processes 75

3.4 Application to the Psoriatic Arthritis Cohort 78

3.5 Discussion and Future Research 84

4 A Two-Phase Model for Chronic Disease Processes Under Intermittent Inspection **89**

4.1 Introduction 89

4.1.1 Disease Processes with Delayed Activity 89

4.1.2 The University of Toronto Psoriatic Arthritis Cohort 92
4.2 Model Formulation and Likelihood under Intermittent Observation 94
4.2.1 General Formulation of a Two-Phase Model 94
4.2.2 Intermittent Assessment and Interval-Censored Data 97
4.3 Piecewise Constant Baseline Functions and the EM Algorithm 98
4.3.1 Complete Data Log-Likelihood 98
4.3.2 The EM Algorithm for Maximum Likelihood Estimation 100
4.3.3 Louis' Method for Estimates Obtained by Simultaneous Maximization 101
4.4 Two-Stage Estimation 104
4.4.1 Description of Two-Stage Procedure 104
4.4.2 Variance Estimation following Two-Stage Estimation 105
4.5 Simulation Studies and Application 107
4.5.1 Design and Interpretation of Simulation Studies 107
4.5.2 Application of Psoriatic Arthritis Data 109
4.6 Discussion 112
Appendix 4.A Evaluation of $Q(\theta;\theta^{(v)})$ for Maximum Likelihood Estimation . . 115
4.A.1 Details of EM Algorithm 115
4.A.2 Evaluations of the Conditional Expectations in the E-Step 118

5 Discussion and Future Research **119**

5.1 Penalized Regression for Interval-Censored Times 119

5.2 Penalized Regression for Truncated and Censored Times 121

5.3 Assessing the Accuracy of Predictive Models with Interval-Censored Data . 124

5.4 Statistical Models for Complex Life History Processes 125

References **128**

List of Tables

2.1 Empirical results for interval-censored data with normally distributed covariates ($p = 100$, $E(X_{ij}) = 0$, $Var(X_{ij}) = 1$ and $\text{corr}(X_{ij}, X_{ik}) = \rho^{|j-k|}$, where $\rho = 0.5$) summarizing the number of correctly (TP) and incorrectly (FP) selected variables along with the median and the standard deviation (SD) of the mean squared error (MSE); P-EM denotes analyses based on the proposed penalized EM method and MID denotes analyses based on a pseudo-dataset obtained by mid-point imputation; the tuning parameter is selected by five-fold cross-validation. 23

2.2 Empirical results for interval-censored data with correlated binary covariates ($p = 100$, $E(X_{ij}) = 0.2$ and $\text{corr}(X_{ij}, X_{ik}) = \rho^{|j-k|}$ if X_{ij}, X_{ik} are in the same block as discussed in Section 2.3 and $\rho = 0.2$) summarizing the number of correctly (TP) and incorrectly (FP) selected variables along with the median and the standard deviation (SD) of the mean squared error (MSE); P-EM denotes analyses based on the proposed penalized EM method and MID denotes analyses based on a pseudo-dataset obtained by mid-point imputation; the tuning parameter is selected by five-fold cross-validation. . 25

2.3 Selected HLA markers and their effects obtained by variable selection with interval-censored data disease progression data in psoriatic arthritis using the LASSO, ALASSO and SCAD penalty functions. 31

2.4 Empirical results for dataset (33.33% left-truncation and 66.67% right-truncation) with binary covariates ($p = 100$, $P(X_{ij} = 1) = 0.5$) summarizing the number of correctly (TP) and incorrectly (FP) selected variables along with the median and the standard deviation (SD) of the mean squared error (MSE). The tuning parameter is selected by AIC, BIC or CV. Sample size $m = 1200$, and $nsim = 100$ simulations. 42

2.A.1 Empirical results for interval-censored data with normally distributed covariates ($p = 10$, $E(X_{ij}) = 0$, $Var(X_{ij}) = 1$ and $\text{corr}(X_{ij}, X_{ik}) = \rho^{|j-k|}$) summarizing the number of correctly (TP) and incorrectly (FP) selected variables along with the median and the standard deviation (SD) of the mean squared error (MSE); P-EM denotes the analyses based on the proposed penalized EM method and MID denotes an analysis based on a pseudo-data set obtained by mid-point imputation; the tuning parameter is selected by five-fold cross validation. 48

2.A.2 Empirical results for interval-censored data with correlated binary covariates ($p = 10$, $E(X_{ij}) = 0.2$ and $\text{corr}(X_{ij}, X_{ik}) = \rho^{|j-k|}$) summarizing the number of correctly (TP) and incorrectly (FP) selected variables along with the median and the standard deviation (SD) of the mean squared error (MSE); P-EM denotes the analyses based on the proposed penalized EM method and MID denotes an analysis based on a pseudo-data set obtained by mid-point imputation; the tuning parameter is selected by five-fold cross validation. 49

2.B.1 Comparison of three methods of choosing tuning parameter: cross-validation (CV), Bayesian information criterion (BIC) and sparse generalized cross-validation (SGCV). Analyses were based on interval-censored responses with correlated binary covariates ($p = 10$) by using proportional hazards models with a piecewise constant baseline hazards with four pieces (PWC-4) and results are summarized in terms of the number of correctly (TP) and incorrectly (FP) selected variables and the median and standard deviation of the mean squared error (MSE). 53

2.B.2 Comparison of three methods of choosing tuning parameter: cross-validation (CV), Bayesian information criterion (BIC) and sparse generalized cross-validation (SGCV). Analyses were based on interval-censored responses with multivariate normal covariates ($p = 100$) by using proportional hazards models with a piecewise constant baseline hazards with four pieces (PWC-4) and results are summarized in terms of the number of correctly (TP) and incorrectly (FP) selected variables and the median and standard deviation of the mean squared error (MSE). 54

2.C.1 Variance estimation by bootstrap for the simulated dataset with multivariate normal covariates and multivariate binary covariates for $\kappa = 1.25$, $\mu = 10$, $\rho = 0.3$. 57

3.1 Empirical performance* of PE; sample size $m = 500$, number of simulations $nsim = 100$. The predictor is $\widehat{Y}(X;\widehat{\theta}) = I(P(T > t_0|X;\widehat{\theta}) > 0.5)$. The covariates have marginal standard normal distributions. The event time T_i follows a Weibull distribution with rate $\kappa\lambda(\lambda t)^{\kappa-1}\exp(X_{i1}\beta_1 + X_{i2}\beta_2)$, where $\beta_1 = \log(2)$, $\beta_2 = \log(1.5)$ and $\kappa = 1.25$. A time homogeneous Poisson process is used for the inspection process with rate $\exp(\gamma_0 + X_{i1}\gamma_1 + X_{i3}\gamma_2)$, where $\gamma_1 = \log(1.1)$ and $\gamma_2 = \log(1.5)$. 76

3.2 Empirical performance* of PE; sample size $m = 500$, number of simulations $nsim = 100$. The predictor is $\widehat{Y}(X;\widehat{\theta}) = I(P(T > t_0|X;\widehat{\theta}) > 0.5)$. The covariates are binary with $P(X_{ij} = 1) = 0.5$, $j = 1, 2, 3$. The event time T_i follows a Weibull distribution with rate $\kappa\lambda(\lambda t)^{\kappa-1}\exp(X_{i1}\beta_1 + X_{i2}\beta_2)$, where $\beta_1 = \log(2)$, $\beta_2 = \log(1.5)$ and $\kappa = 1.25$. A time homogeneous Poisson process is used for the inspection process with rate $\exp(\gamma_0 + X_{i1}\gamma_1 + X_{i3}\gamma_2)$, where $\gamma_1 = \log(2)$ and $\gamma_2 = \log(2.5)$. 77

3.3 Empirical performance* of PE; sample size $m = 500$, number of simulations $nsim = 100$. The inspection process is generated by a renewal process. The gap times between two consecutive inspections are generated by a Gamma distribution with shape η and rate $\exp(\gamma_0 + X_{i1}\gamma_1 + X_{i3}\gamma_2)$, where $\gamma_1 = \log(1.1)$, $\gamma_2 = \log(1.5)$ and $\eta = 1.25$, 1.5 and 2. (Normal Cases, $X_{i2} \perp X_{i3}$) 79

3.4 Empirical performance* of PE; sample size $m = 500$, number of simulations $nsim = 100$. The inspection process is generated by a renewal process. The gap times between two consecutive inspections are generated by a Gamma distribution with shape η and rate $\exp(\gamma_0 + X_{i1}\gamma_1 + X_{i3}\gamma_2)$, where $\gamma_1 = \log(1.1)$, $\gamma_2 = \log(1.5)$ and $\eta = 1.25$, 1.5 and 2. (Normal Cases, $X_{i2} \not\perp X_{i3}$) 80

3.5 Empirical performance* of PE; sample size $m = 500$, number of simulations $nsim = 100$. The inspection process is generated by a renewal process. The gap times between two consecutive inspections are generated by a Gamma distribution with shape η and rate $\exp(\gamma_0 + X_{i1}\gamma_1 + X_{i3}\gamma_2)$, where $\gamma_1 = \log(2)$, $\gamma_2 = \log(2.5)$ and $\eta = 1.25$, 1.5 and 2. (Binary Cases, $X_{i2} \perp X_{i3}$) . 81

3.6 Empirical performance* of PE; sample size $m = 500$, number of simulations $nsim = 100$. The inspection process is generated by a renewal process. The gap times between two consecutive inspections are generated by a Gamma distribution with shape η and rate $\exp(\gamma_0 + X_{i1}\gamma_1 + X_{i3}\gamma_2)$, where $\gamma_1 = \log(2)$, $\gamma_2 = \log(2.5)$ and $\eta = 1.25$, 1.5 and 2. (Binary Cases, $X_{i2} \not\perp X_{i3}$) . 82

4.1 Empirical performance† of estimators; sample size $m = 500$, number of simulations $nsim = 500$, $\alpha_1 = 0.036$, $\alpha_2 = 0.5$, $\beta_1 = (0.5, 0.5)$, $\beta_2 = (-0.5, -0.5)$; ASE are average of standard errors estimated via methods in Section 4.3.3 (EM-MLE) and Section 4.4.2 (EM-TS). 108

4.2 Results of fitting piecewise constant baseline hazard model for the duration of the indolent period and piecewise constant baseline rate model for the occurrence of joint damages under simultaneous (EM-MLE) and two-stage (EM-TW) estimation; $p-$values are based on Wald tests. 111

4.A.1 Pseudo-dataframe for the maximization of Q_1. 116

4.A.2 Pseudo-dataframe for the maximization of Q_2. 117

List of Figures

1.1 Lexis diagram of event and assessment times on the scale of calendar time (horizontal axis) and the time since initiating event (the vertical axis). . . 3

2.1 Plots of the estimated cumulative distribution functions for the time from psoriatic arthritis diagnosis and clinic entry (Kaplan-Meier estimate) and the times between radiological assessments based on a semi-Markov model with a gamma frailty (panel (a)) and the Turnbull estimate with a pointwise 95% confidence band for the marginal cumulative distribution function of the time from disease onset to arthritis mutilans (panel (b)). 13

2.2 Box plots of the error for the estimated regression coefficients $\widehat{\beta}_k - \beta_k$, $k = 5, 22, 95, 96$, for each penalty function for datasets with correlated binary covariates ($p = 100$) with $\kappa = 1.25$, $\mu = 20$. 27

2.3 Plots of the cross validation statistics and shrinkage of coefficients in the PsA dataset based on piecewise constant hazard model via EM algorithm with the LASSO, ALASSO and SCAD penalty functions. 32

2.4 Lexis diagrams of the calendar times of birth (B), onset of psoriasis (E_0) and onset of psoriatic arthritis (E_1), along with screening times (A_0) for UTPC (left panel) and the UTPAC (right panel). 36

2.5 Plots of the BIC (1st column), 5-fold cross-validation statistic (2nd column) and shrinkage estimates of coefficients (3rd column) against the tuning parameter from penalized regression of the PsA dataset based on a piecewise constant hazard model (PWC-4) fitted via an EM algorithm with the LASSO, ALASSO or SCAD penalty. The fixed covariates are gender and the onset age of psoriasis. 44

2.A.1 Box plots of the error for the estimated regression coefficients $\widehat{\beta}_k - \beta_k$, $k = 1, 2, 3, 5$, for each penalty function for datasets with correlated binary covariates ($p = 10$) with $\kappa = 1.25$, $\mu = 10$, $\rho = 0.3$. 50

3.1 A multistate diagram for joint consideration of event, random drop-out and assessment times. 67

3.2 Four schematic diagram enumerating possible combinations of (Y, Δ); the solid lines denote observations in which the event status is known at t_0 and the dashed lines denote individual whose event status cannot be classified and hence who are excluded from the sum in (3.13); the solid dots denote the (unobserved) exact event times. 68

3.3 Plots of the estimates of the prediction error against t_0 with a binary predictor $I(P(T > t_0) > 0.5)$ for the models obtained from penalized regression with the LASSO, ALASSO and SCAD penalty functions. The upper panels show estimates obtained by using unweighted, inverse probability weighted and augmented inverse weighted methods; the bottom panels show estimates obtained by using model- and imputation-based methods. 86

3.4 Plots of the estimates of the AUC against t_0 with a binary predictor $I(P(T > t_0) > c)$, where c ranging from 0 to 1. The response models are obtained from penalized regression with the LASSO, ALASSO and SCAD penalty functions. 87

3.5 Estimates of ROC curves at t_0 =10, 20 and 30 years by inverse probability weighted (upper panels) and augmented inverse probability (bottom panels) methods for the models obtained from penalized regression with the LASSO, ALASSO and SCAD penalties with the tuning parameter selected by the 5-fold CV. 88

4.1 Plot of assessment times (hatch marks) and time of radiological damaged joints detected between assessments (solid points) from onset of psoriatic arthritis for a selected sample of patients from University of Toronto Psoriatic Arthritis Clinic. The dashed line denotes time from disease onset to first occurrence of joint damage, and the solid line denotes the period of disease progression following onset of damage. 93

4.2 Lexis diagram of event and assessment times on the scale of disease duration (t) on the horizontal axis and the time since start of phase II (t^*) on the vertical axis. 95

4.3 Estimates of the probability of having at least one damaged joint as measured from the time since disease onset (left panel) and the expected number of damaged joints for the time since disease onset (right panel). 112

5.1 Multistate diagram illustrating the course of disease in psoriatic arthritis cohort. 122

5.2 Multistate diagram illustrating the two phases of the disease process. . . . 126

Chapter 1

Introduction

1.1 Motivating Research Program

1.1.1 Overview

This research is directed at the development of innovative statistical models and methods to address challenging problems arising in research at the Centre for Prognosis Studies in Rheumatic Diseases at the Toronto Western Hospital. This centre created the University of Toronto Psoriasis Clinic (UTPC) in 2008 to study the course of psoriasis (Ps), a chronic inflammatory skin condition which affects up to 3% of the population (Schäfer, 2006). Screened patients identified as having psoriasis are recruited to this clinical registry and upon entry they undergo a detailed clinical examination, provide samples for genetic testing, are then followed prospectively according to a standardized protocol; clinical assessments are planned to take place every 6-12 months, but ultimately there is considerable variation in the times of the follow-up assessments.

Approximately 30% of psoriasis patients develop psoriatic arthritis (PsA), a rheumatological disorder featuring inflammatory psoriatic disease as well as inflammation and damage in and around the joints of several areas including the wrists, hands, knees, ankles, lower back, and neck (Chandran et al., 2010). The University of Toronto Psoriatic Arthritis Clinic (UTPAC) was launched in 1977 to study this complex disease (Gladman et al., 2008). While patients are recruited to this registry for a variety of reasons, a primary method is through the use of a population-based screening tool in the form of a 10 item questionnaire (Tom et al., 2015). Individuals suspected of having psoriatic arthritis based on this tool are invited to attend the clinic for a more definitive diagnosis, and those found to have the disease are invited to join the UTPAC. Upon entry to UTPAC, as in the UTPC, a detailed history is taken, patients undergo a thorough clinical and radiological examination, and samples are collected for genetic testing. The genotypes of HLA-A, HLA-B, HLA-C, HLA-DR and HLA-DQ alleles were collected in both cohorts. Patients are then scheduled to undergo detailed annual clinical examinations and biannual radiological examinations.

1.1.2 Psoriasis and Psoriatic Arthritis

Patients with psoriasis which is uncomplicated with arthritis (PsC) at the screening time were recruited into the University of Toronto Psoriasis Cohort. To date, it has enrolled several hundred patients, but the analyses reported here are based on data from 580 patients in the registry as of 2013; there are 250 women and 330 men. The age at clinic entry has a mean of 46.4 and standard deviation 13.5 years. The median time from clinic entry to the last visit is 1.2 years with maximum of 7.1 years, and the number of visits ranges from 1 to 8 with a median of 2 visits. And the mean (sd) of age of diagnosis of Ps is 30.4

(16.2) years.

The University of Toronto Psoriatic Arthritis cohort recruited patients with psoriatic arthritis at the screening time, it has over 1000 patients so far. Among the 1215 patients in the registry as of 2013, there are 524 women and 691 men. The mean age at clinic entry is 44.1 years with a standard deviation of 13.0 years. The median time from clinic entry to the last visit is 4.9 years with maximum of 36.7 years, and the number of visits ranges from 1 to 57 with a median of 6 visits. The mean (sd) of age of diagnosis of Ps and PsA is 28.6 (14.7) and 37.2 (13.4) years respectively.

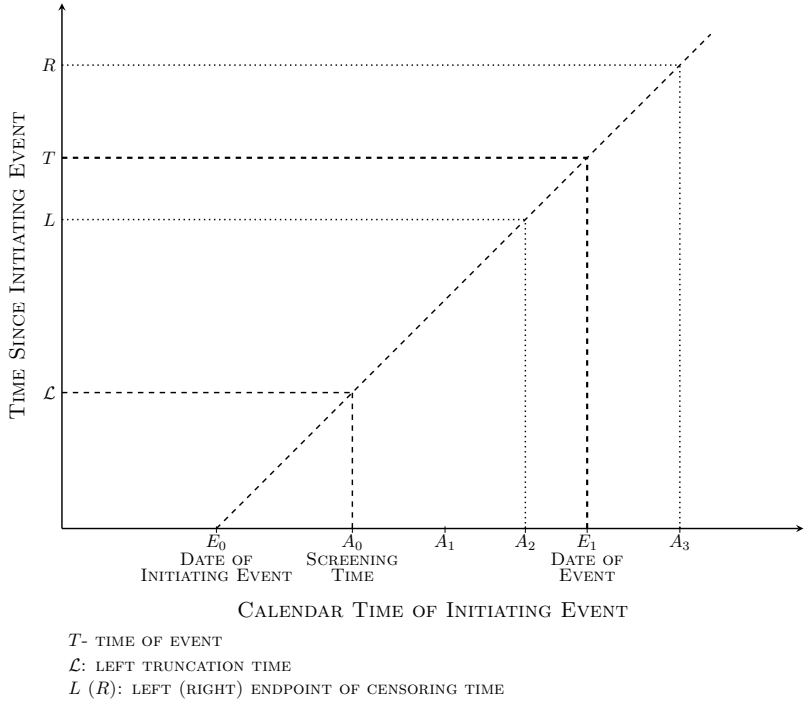


Figure 1.1: Lexis diagram of event and assessment times on the scale of calendar time (horizontal axis) and the time since initiating event (the vertical axis).

Identification of genetic factors putting psoriasis patients at elevated risk of PsA is important as it will enable high risk psoriasis patients to be monitored more closely to

ensure treatments geared toward the prevention of damage from arthritis are administered in a timely fashion. Such predictive models can also help guide the selection of high risk patients for inclusion in clinical trials of experimental prophylactic treatments. With the increasing availability of large prospective disease registries, scientists studying the course of chronic conditions often have access to multiple data sources, with each source generated based on its own entry conditions. The different entry conditions of the various registries may be explicitly based on the response process of interest, in which case the statistical analysis must recognize the unique truncation schemes. Moreover, intermittent assessment of individuals in the registries can lead to interval-censored times of interest.

Figure 1.1 is a Lexis diagram we introduce to illustrate the point that in the Ps cohort, the time from Ps onset to PsA onset is subject to left-truncation and interval-censoring. In this diagram the onset of Ps is the initiating event and the onset of PsA is the event of interest. The date of screening and recruitment to the Ps registry is denoted by A_0, and the follow-up assessments are denoted by A_1, A_2 and A_3. Since an individual will only be recruited to the Ps registry if they are PsA-free then the time $\mathcal{L} = A_0 - E_0$ is the left-truncation time for the time $T = E_1 - E_0$ of interest. The last PsA-free assessment and the first assessment after the onset of PsA define the left (L) and right (R) endpoints of the censoring interval respectively. These latter times are depicted on the vertical axis of Figure 1.1. For patients in the PsA registry there are analogous right-truncation conditions which must be addressed. We develop, evaluate and apply methods for variable selection with left- and right-truncated data with event times subject to interval-censoring.

1.1.3 Arthritis Mutilans

Psoriatic arthritis can be classified into 5 distinct sub-types according to the phenotypic presentations. Arthritis mutilans is considered to be the most severe form of PsA in which patients experience deformity and severe destruction of the joints. While there is no clinical agreement on how to precisely define arthritis mutilans, it represents a state of significant joint damage arising from an extreme form of arthritic component of the disease; here we define it as present if an individual has 5 or more joints with the advanced stage of damage according to the modified Steinbrocker score. It is important to identify clinical predictors and biomarkers for arthritis mutalins in order to prevent further joint destruction. Data from 604 patients in the PsA registry are used in these analyses. A total of 96 HLA markers were used in the study, but 20 of these markers had a frequency in the sample of less than 1% and so were excluded from further consideration, leaving 76 markers to select from. We expand on the description of the problems and approaches in the following subsections.

1.2 General Introduction to Research Topics

The previous section gave a brief overview of the disease processes of interest and the data available for analysis. The following is a similarly brief overview of the types of methodological problems to be considered in the future chapters. A literature review, notation and other details are contained in the respective chapters, which for the most part are self-contained.

1.2.1 Penalized Regression for Interval-Censored Times of Disease Progression

In the context of time to event analysis, much of the methodological work on variable selection has been carried out for right-censored data. When events of interest are only detectable upon imaging, as is the case for joint damage which is assessed radiologically at periodic examination times, times of joint damage are interval-censored. Truncation is also often a factor in life history analysis when using data from different registries or other sources, and truncation conditions must be recognized for analyses to be valid.

We propose an algorithm for penalized regression (e.g. LASSO, adaptive LASSO and SCAD) to handle truncated and interval-censored times in Chapter 2. A flexible parametric model with piecewise constant baseline hazard function is adopted and an expectation-maximization algorithm is described which is empirically shown to perform well. The developments are presented in two stages motivated by two distinct problems. We first develop an algorithm dealing with interval-censored responses with a view to identifying markers associated with the development of arthritis mutilans in patients with PsA. Several important human leukocyte antigen (HLA) variables are identified for further investigation.

An extension is then described and evaluated to deal with truncated data using a a penalized Turnbull-type complete data likelihood which incorporates information from individuals who did not satisfy the selection criteria. Simulation studies demonstrate good empirical performance and an application to the motivating study identifies HLA markers associated with the onset of psoriatic arthritis in patients with psoriasis from both the Ps and PsA cohorts. Both left- and right-truncation must be dealt with in analyses using data from both cohorts.

1.2.2 Assessing the Accuracy of Predictive Models with Interval-Censored Data

Assessing the statistical performance of a prediction model is important for establishing the validity of a prognostic model and hence in directing medical and clinical decisions. There has been a lot of work directed at the evaluation of predictive models with right-censored data. Methods include estimation of overall prediction error by using loss functions, and evaluation of discriminative ability through use of the receiver operating characteristic (ROC) curves for event status.

In Chapter 3, inverse probability weighted (IPW) and augmented inverse probability weighted (AIPW) estimators are developed and evaluated based on the mean prediction error and the area under the receiver operating characteristic curve to evaluate the performance of predictive models for interval-censored response. The weights are estimated through the use of a multistate model facilitating the joint consideration of the event, inspection and drop-out processes. We empirically investigate the performance of the proposed methods and illustrate their application in the context of a motivating rheumatology study in which HLA markers are used for predicting disease progression in psoriatic arthritis patients.

1.2.3 A Two-Phase Model for Chronic Disease Processes Under Intermittent Inspection

In many chronic diseases, there is often considerable variation in the course of disease. In rheumatoid arthritis, for example, the length of inactive disease may vary extensively between individuals. Moreover, once the disease becomes "active", some individuals expe-

rience rapid progression while others experience minimal disease activity. One approach to deal with this kind of situation is using models with multiple components. In Chapter 4, we propose a two-phase model which has one component for the time from disease occurrence to the onset of damage, and another component which characterizes the nature of the damage process once it begins. This model can be used to separately examine prognostic factors for the length of the inactive phase as well as factors prognostic for the nature and rate of damage in the active phase. It can therefore be used to obtain a more appropriate representation of a complex multi-phase disease process, can help identify different types of risk factors, and could yield more accurate prediction models.

With interval-censored recurrent event data, all we known are counts of the occurrences of events between assessments. Therefore, the two-phase model involves modeling interval-censored times of the precipitating event (i.e. the first joint to become damaged) and panel count data with a latent time origin. Simulation studies are conducted to examine the performance of the proposed approach to model fitting. Application to data from the motivating study of disease progression in psoriatic arthritis is also given.

Chapter 2

Penalized Regression for Interval-Censored Times of Disease Progression

2.1 Introduction

2.1.1 Variable Selection and Penalized Regression

The literature on statistical methods for variable selection has developed considerably over the last twenty years. Breiman (1996) pointed out that the traditional method of best subset selection was unstable and that this instability could lead to poor performance regarding prediction. While ridge regression (Hoerl and Kennard, 1970) imposes some shrinkage which leads to more stable models, it does not set any coefficients to zero and therefore does not "select" key variables. Tibshirani (1996) proposed a "least absolute

shrinkage and selection operator", widely known by the acronym "LASSO". The LASSO attempts to maintain the advantages of both subset selection and ridge regression by shrinking some coefficients and setting other coefficients to zero through the addition of a particular penalty function to the log-likelihood. Several other penalty functions have been developed and studied over the last decade to cope with high dimensional predictor spaces, including the smoothly-clipped absolute deviation (SCAD) (Fan and Li, 2001; Zou and Li, 2008), the adaptive LASSO (Zou, 2006), the elastic net (Zou and Hastie, 2005), the grouped LASSO (Yuan and Lin, 2005), the fused LASSO (Tibshirani et al., 2005) and the minimax concave penalty (MCP) (Zhang, 2010).

While much of the work on variable selection techniques was initially carried out in the context of continuous responses (Tibshirani, 1996; Fan and Li, 2001; Zou, 2006), advances have been made to deal with binary responses (Park and Hastie, 2007; Friedman et al., 2010) and time to event responses (Tibshirani, 1997; Fan and Li, 2002; Zhang and Lu, 2007). For the latter, the penalty term is typically applied to the partial likelihood (Cox, 1975) arising from a semiparametric Cox regression model (Cox, 1972) when data are right-censored.

Witten and Tibshirani (2009) give an excellent overview of the challenges arising with particularly high dimensional covariate data in settings with censored outcomes and provide an extensive discussion of the specific objectives one might have in particular scientific contexts; another useful account can be found in Li and Ma (2013). The inherent difficulty in obtaining robust and generalizable findings from samples with censored responses and high dimensional covariates is evident from the inconsistency of findings across seemingly similar patient populations and the modest gains that have been made despite considerable advances in biotechnology and statistical methods (McShane et al., 2005a). The limitations analysts face due to inadequate sample size of individual studies (Polley et al., 2013) and

the inconsistency of findings across studies has led to an increased interest in synthesizing findings over multiple studies. Assimilating information from several sources can be helpful, but it is important to clearly understand the differences between the frameworks and goals of the studies contributing to this synthesis. Guidelines have been developed for reporting findings from biomarker studies with this in mind, which advocate clear statements of study objectives, study design, methods of processing samples, and the approach to statistical analysis (McShane et al., 2005b; Altman et al., 2012).

Many prospective studies, however, have the added complication that the event times of interest are subject to interval-censoring. In clinical trials involving cancer patients at risk of metastases, for example, new lesions are only detectable when imaging assessments are carried out (Hortobagyi et al., 1996), and the precise time from randomization to the development of a new lesion is unavailable. In patients infected with cytomegalovirus, the time from infection to viral shedding in the blood is only known to lie between the last negative and first positive serum sample (Betensky and Finkelstein, 1999; Cook et al., 2008). Vertebral fractures in patients with osteoporosis are often asymptomatic, and their occurrence is only detected upon a radiographic examination yielding evidence of a new fracture (Riggs et al., 1990). Sun (2006) gives an excellent account of statistical methods for parametric and semiparametric analysis of interval-censored failure time data.

We consider the problem of variable selection in the context of interval-censored time to event data. We adopt a flexible piecewise exponential model (Friedman, 1982) for the event of interest and penalize the complete data likelihood constructed by treating the interval-censored failure times as known. An expectation-maximization (EM) algorithm (Dempster et al., 1977) is then used for variable selection through optimization of the penalized observed data likelihood. The LASSO, adaptive LASSO and SCAD penalty functions are considered.

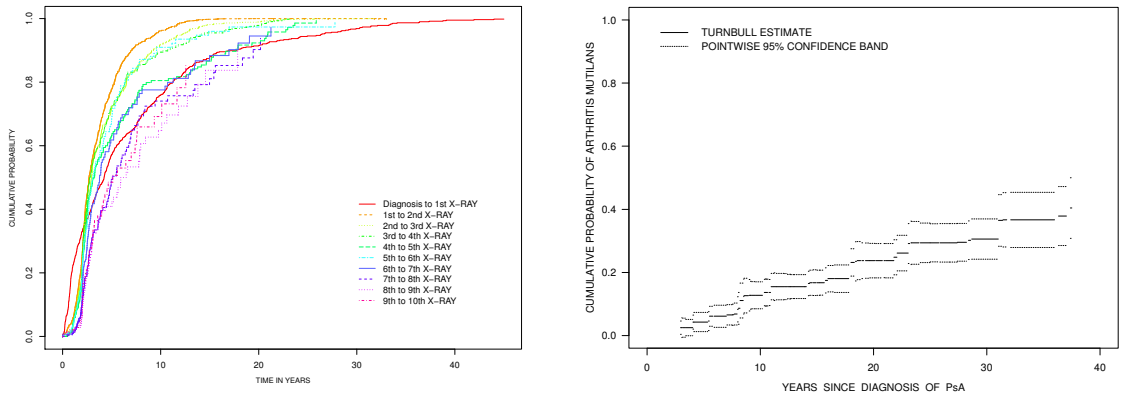
The remainder of this chapter is organized as follows. In Section 2.1.2 we describe the motivating study with the goal of identifying key human leukocyte antigens associated with the development of arthritis mutilans in a cohort of individuals with psoriatic arthritis. In Section 2.2 we describe a penalized expectation-maximization algorithm based on a piecewise exponential response model, for which existing techniques for variable selection can be exploited to handle interval-censored event times. This is the primary contribution of this chapter. Simulation studies involving multivariate normal covariates vectors are reported in Section 2.3, which demonstrate superior performance of the proposed method over analyses based on mid-point imputation (Lindsey and Ryan, 1998). Additional simulation studies for correlated binary covariates are described in *Appendix 2.A* and studies of different criteria for selection of tuning parameters (Bradic et al., 2011) are given in *Appendix 2.B*. The data from the psoriatic arthritis clinic are analyzed in Section 2.4 using a variety of penalty functions, and concluding remarks are given in Section 2.5. An extension of this algorithm is described and evaluated to deal with truncated data in Section 2.6.

2.1.2 Prognostic HLA Markers in Psoriatic Arthritis

The University of Toronto Psoriatic Arthritis Clinic is a tertiary referral center for individuals with psoriatic arthritis (PsA), an immunological condition which features both skin and joint involvement (Chandran et al., 2010). A registry was created in 1976, which has been recruiting and following patients continuously since its inception, making it one of the largest cohorts of patients with PsA in the world.

Patients undergo a detailed clinical and radiological examination upon entry to the clinic, and provide serum samples for genetic testing. Follow-up clinical and radiological assessments are scheduled annually and every two years respectively in order to track

changes in joint damage. At each radiological assessment the degree of damage is recorded in sixty-four joints on a five-point scale (Rahman et al., 1998). Arthritis mutilans is a particularly aggressive form of arthritis characterized here by five or more joints with the highest grade of damage. Identification of genetic features associated with this aggressive form of arthritis is important to help identify patients warranting prophylactic treatment with more effective but costly anti-TNF therapy (Kyle et al., 2005) and to help guide the selection of high risk patients for inclusion in clinical trials of experimental treatments. The aim of the current analysis is to identify key human leukocyte antigens which are associated with increased risk of arthritis mutilans in this cohort of patients.



(a) (b)

Figure 2.1: Plots of the estimated cumulative distribution functions for the time from psoriatic arthritis diagnosis and clinic entry (Kaplan-Meier estimate) and the times between radiological assessments based on a semi-Markov model with a gamma frailty (panel (a)) and the Turnbull estimate with a pointwise 95% confidence band for the marginal cumulative distribution function of the time from disease onset to arthritis mutilans (panel (b)).

To date, 1191 patients have been recruited to the University of Toronto Psoriatic Arthritis Clinic, and 604 of these have undergone genetic testing to determine their human leukocyte antigen profile. A total of 96 human leukocyte antigen covariates were available for study but 20 of these markers had a frequency in the sample of less than 1% and so were excluded from further consideration. Among the 604 patients the median time from clinic entry to last radiological assessment is 6.3 years and there is a median of 3 radiological assessments per patient. To give a sense of the variability in the times between radiological assessments, the estimated cumulative distribution functions of the times between the first 10 radiological assessments are displayed in Figure 2.1a. To account for the between individual variation in the propensity to attend the clinic (or equivalently, to account for the within-individual dependence in the gap times), the estimated cumulative distribution functions were obtained by fitting a semi-Markov model stratified on the cumulative number of radiological assessments and with an individual-specific gamma distributed frailty term (Klein, 1992; Nielsen et al., 1992). The median inter-assessment times range from 2.7 years for the first two or three assessments after clinic entry, to over 6 years for later assessments. Also plotted is a marginal Kaplan-Meier estimate of the time from diagnosis of psoriatic arthritis to clinic entry; the median of this distribution is roughly similar to the median times between assessments but there are more observations in the right tail of this distribution.

Five hundred and seven (83.9%) of the 604 individuals in this dataset were not observed to develop arthritis mutilans and hence provided right-censored times, whereas 97 (16.1%) individuals were known to develop arthritis mutilans and so yielded interval-censored times. Figure 2.1b contains a nonparametric estimate of the cumulative distribution function for the time from disease onset to arthritis mutilans based on the Turnbull algorithm (Turnbull, 1976) along with pointwise 95% confidence bands. The estimate reflects a

steadily increasing risk with roughly 23% developing the condition within 20 years of disease onset.

2.2 Variable Selection with Interval-Censored Data

2.2.1 Notation and the Penalized Complete Data Likelihood

Here we consider the problem of variable selection with interval-censored data. In many settings, including the motivating study, a natural time origin is the time of disease onset. We let T_i denote the time from disease onset to the event of interest for individual i in a sample of m independent individuals, $i = 1, \ldots, m$. We assume individuals are examined at assessment times governed by an independent inspection process (Grüger et al., 1991; Cook and Lawless, 2014) and let $\mathcal{C}_i = [L_i, R_i)$ denote the interval known to contain the event for subject i, $i = 1, \ldots, m$. For left-censored data $L_i = 0$, for right-censored data $R_i = \infty$, and for interval-censored data $0 < L_i < R_i < \infty$. We let $X_i = (X_{i1}, \ldots, X_{ip})'$ denote a $p \times 1$ covariate vector.

We wish to examine the relation between the covariates and the time of interest based on a proportional hazards model with $h(t|X_i;\theta) = h_0(t;\alpha)\exp(X_i'\beta)$ where α parameterizes the baseline hazard, $\beta = (\beta_1, \ldots, \beta_p)'$, and $\theta = (\alpha', \beta')'$. We adopt a weakly parametric piecewise constant baseline hazard function which requires specification of the number and location of times the hazard changes value; we subsequently refer to these as break-points. If $0 = b_0 < b_1 < \cdots < b_{K-1} < b_K = \infty$ denote K break-points, the baseline hazard function is $h_0(s;\alpha) = \exp(\alpha_k)$, for $s \in \mathcal{B}_k = [b_{k-1}, b_k)$, $k = 1, \ldots, K$. The survivor function is then $\mathcal{F}(t|X_i;\theta) = \exp\{-H(t|X_i;\theta)\}$ where $H(t|X_i;\theta) = \int_0^t h(s|X_i;\theta)ds$. Given the covariate

vector X_i and an independent inspection process, the observed (partial) likelihood is

$$L(\theta) \propto \prod_{i=1}^{m} \{\mathcal{F}(L_i|X_i;\theta) - \mathcal{F}(R_i|X_i;\theta)\}$$

and the corresponding observed data log-likelihood is

$$\log L(\theta) \propto \sum_{i=1}^{m} \log \{\mathcal{F}(L_i|X_i;\theta) - \mathcal{F}(R_i|X_i;\theta)\} . \tag{2.1}$$

When viewing this as a variable selection problem, we are specifically interested in identifying the component covariates for which the respective regression coefficients are non-zero. Many common methods of variable selection are based on a penalized likelihood of the form

$$\log L_{\text{PEN}}(\theta) = \frac{1}{m} \log L(\theta) - p_{\gamma,\lambda}(\beta) , \tag{2.2}$$

where the function $p_{\gamma,\lambda}(\beta)$ determines the extent of the penalty for each value of β, modulated by the tuning parameters (γ, λ). Ridge regression (Hoerl and Kennard, 1970) is implemented with the L_2 penalty $p_{\gamma,\lambda}(\beta) = \lambda \sum_{j=1}^{p} \beta_j^2$ and the LASSO (Tibshirani, 1996) uses the L_1 penalty $p_{\gamma,\lambda}(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|$; there is no tuning parameter γ in these penalty functions. The value of the scalar λ is typically found by cross-validation (Shao, 1993) or generalized cross-validation (Golub et al., 1979). The adaptive LASSO uses adaptively weighted L_1 penalties of the form

$$p_{\gamma,\lambda}(\beta) = \sum_{j=1}^{p} \lambda_j |\beta_j| , \tag{2.3}$$

with small penalties λ_j chosen for large coefficients to reduce their shrinkage, and large penalties for small coefficients to address the selection objective (Zou, 2006). One option is to set $\lambda_j = \lambda/|\widetilde{\beta}_j|$, where $\widetilde{\beta} = (\widetilde{\beta}_1, \widetilde{\beta}_2, \ldots, \widetilde{\beta}_p)'$ is the maximum likelihood estimate (Zou, 2006; Zhang and Lu, 2007). Alternatively, the penalties can be updated iteratively. In this

case, at the $(\ell+1)$st implementation, λ_j is set to $\lambda_j^{(\ell)} = \lambda/|\widetilde{\beta}_j^{(\ell)}|$ where $\widetilde{\beta}^{(\ell)}$ is obtained on the ℓth iteration; when $\ell = 0$, we set $\lambda_j^{(0)} = \lambda/|\widetilde{\beta}_j|$ as in the first implementation (Fan and Lv, 2010). We investigate the iterative implementation of the adaptive LASSO in the next section.

The smoothly clipped absolute deviation (SCAD) penalty proposed by Fan and Li (2001) is defined by

$$p'_{\gamma,\lambda}(\beta) = \lambda \sum_{j=1}^{p} \left\{ I(|\beta_j| \leq \lambda) + \frac{(\gamma\lambda - |\beta_j|)_+}{(\gamma-1)\lambda} I(|\beta_j| > \lambda) \right\} ,$$

where $\gamma > 2$ and $y_+ = I(y \geq 0) \times y$. This penalty function is continuously differentiable on $(-\infty, 0) \cup (0, \infty)$, but singular at 0 with its derivatives zero outside the range $[-\gamma\lambda, \gamma\lambda]$. Therefore, the SCAD penalty results in "small" coefficients being set to zero, "moderate" coefficients being shrunk towards zero, and "large" coefficients retained as they are. In principle, the optimal pair (γ, λ) could be obtained using a two-dimensional grid search by cross validation or generalized cross validation. From empirical work, Fan and Li (2001) suggest $\gamma = 3.7$ is a reasonable choice for a variety of problems and we use this in what follows and select λ by (generalized) cross validation.

2.2.2 An Expectation-Maximization Algorithm

We develop here an expectation-maximization algorithm for optimizing (2.2) using available algorithms for penalized regression (Dempster et al., 1977). We do this by considering a complete data likelihood in which the latent event time is treated as known rather than interval-censored.

Let $D_k(u) = I(u \in \mathcal{B}_k)$ denote whether or not the time u is in the interval $\mathcal{B}_k$ and $W_k(u) = \int_0^u I_k(s)ds$ denote the duration at risk in interval k over $[0, u)$. If the event time

t_i were known, then under the piecewise constant model and given a covariate vector X_i, the complete data log-likelihood $\log L_{\text{COMP}}(\theta)$ would be

$$\sum_{i=1}^{m}\sum_{k=1}^{K}\left[D_k(t_i)\left\{\log(\rho_k)+X_i'\beta\right\}-W_k(t_i)\rho_k\exp(X_i'\beta)\right]\ . \tag{2.4}$$

Let $Z_{ik\ell}=I(k=\ell)$ indicate $k=\ell$, $\ell=1,\ldots,K$ and let $Z_{ik}=(Z_{ik1},\ldots,Z_{ikK})'$ denote the corresponding vector of these indicator functions, $k=1,\ldots,K$; thus $Z_{i1}=(1,0,\ldots,0)'$, $Z_{i2}=(0,1,\ldots,0)'$, …, $Z_{ik}=(0,0,\ldots,1)'$. If $\alpha_k=\log\rho_k$ for $k=1,\ldots,K$, and $\alpha=(\alpha_1,\ldots,\alpha_K)'$, we can write

$$\log L_{\text{COMP}}(\theta)=\sum_{i=1}^{m}\sum_{k=1}^{K}\left\{D_k(t_i)V_{ik}'\theta-W_k(t_i)\exp(V_{ik}'\theta)\right\}\ . \tag{2.5}$$

where $V_{ik}=(Z_{ik}',X_i')'$ and $\theta=(\alpha',\beta')'$. Since the penalty in (2.2) is simply a function of the regression parameters, maximization of the penalized likelihood (2.2) can be achieved by applying the EM algorithm to the penalized complete data likelihood

$$\frac{1}{m}\log L_{\text{COMP}}(\theta)-p_{\gamma,\lambda}(\beta)\ . \tag{2.6}$$

The expectation-maximization algorithm proceeds as follows:

The E-step

We let $D_i=(L_i,R_i,X_i)$ represent the observed data from individual i and $D=\{D_i,i=1,\ldots,m\}$ denote the observed data for the full sample. The conditional expectation of (2.6) at the $(r+1)$st iteration is evaluated as

$$Q_{\text{PEN}}(\theta;\theta^{(r)})=E\left\{\log L_{\text{COMP}}(\theta)|D;\theta^{(r)}\right\}-p_{\gamma,\lambda}(\beta)\ , \tag{2.7}$$

where $\theta^{(r)}$ is the estimate obtained from the rth iteration. The required conditional expectations are therefore $\widehat{\Delta}_{ik}^{(r)}=E[D_k(T_i)|D_i;\theta^{(r)}]$ and $\widehat{\omega}_{ik}^{(r)}=E[W_k(T_i)|D_i;\theta^{(r)}]$.

Let $\mathcal{C}_{ik} = \mathcal{C}_i \cap \mathcal{B}_k = [L_{ik}, R_{ik})$ denote the sub-interval of the censoring interval $\mathcal{C}_i$ contained within $\mathcal{B}_k$. When $\mathcal{C}_{ik} = \emptyset$, the required expectations are relatively easy to compute since, for instance, it is clear that $D_k(t_i) = 0$ and $\widehat{\Delta}_{ik}^{(r)} = 0$. Moreover, if $b_k < L_i$, then it is known that individual i was at risk for the entire interval $\mathcal{B}_k$ so $W_k(t_i) = \widehat{\omega}_{ik}^{(r)} = b_k - b_{k-1}$, and if $R_i < b_{k-1}$, then $W_k(t_i) = \widehat{\omega}_{ik}^{(r)} = 0$ since they are known to have failed prior to the start of interval $\mathcal{B}_k$. If, on the other hand, $\mathcal{C}_{ik} \neq \emptyset$ then we have:

$$\widehat{\Delta}_{ik}^{(r)} = \frac{\mathcal{F}(L_{ik}|X_i;\theta^{(r)}) - \mathcal{F}(R_{ik}|X_i;\theta^{(r)})}{\mathcal{F}(L_i|X_i;\theta^{(r)}) - \mathcal{F}(R_i|X_i;\theta^{(r)})} \tag{2.8}$$

$$\widehat{\omega}_{ik}^{(r)} = \max(L_i - b_{k-1}, 0) + \int_{\max(L_i, b_{k-1})}^{\min(R_i, b_k)} \frac{\mathcal{F}(s|X_i;\theta^{(r)})}{\mathcal{F}(L_i|X_i;\theta^{(r)}) - \mathcal{F}(R_i|X_i;\theta^{(r)})} \mathrm{d}s \; . \tag{2.9}$$

Given these results, (2.7) can be written more explicitly as

$$\sum_{i=1}^{m} \sum_{k=1}^{K} \left\{ \widehat{\Delta}_{ik}^{(r)} V_{ik}'\theta - \widehat{\omega}_{ik}^{(r)} \exp(V_{ik}'\theta) \right\} - p_{\gamma,\lambda}(\beta) \; . \tag{2.10}$$

The M-step

The objective function (2.10) has the form of a penalized Poisson likelihood (McCullagh and Nelder, 1989). The value $\theta^{(r+1)}$ that maximizes (2.10) can therefore be obtained using software for penalized Poisson regression by creating a dataset comprised of pseudo-individuals indexed by (i, k). If $R_i \geq b_{k-1}$, then at the $(r+1)$st iteration this dataset should include a contribution from pseudo-individual (i, k) with pseudo-count $\widehat{\Delta}_{ik}^{(r)}$ and offset $\log \widehat{\omega}_{ik}^{(r)}$; if $R_i < b_{k-1}$ then no such contribution is required. The function $Q_{\mathrm{PEN}}(\theta; \theta^{(r)})$ is then maximized with respect to θ using standard software for penalized Poisson regression (e.g. the `glmnet(.)` function (R Core Team, 2013; Friedman et al., 2010) or `SIS(.)` (Fan et al., 2010)).

This optimization procedure is repeated iteratively with updated values of (2.8) and (2.9) in (2.10) until the difference between successive estimates becomes small enough to satisfy convergence criterion. In our implementation the iterations were terminated when

$$\max_j(|\theta_j^{(r+1)} - \theta_j^{(r)}|/|\theta_j^{(r)}|) < \epsilon \ ,$$

where $\epsilon = 10^{-6}$.

Selection of the Optimal Tuning Parameter λ_{OPT}

The criterion for selecting the optimal λ is similar to the traditional cross validation. Here we use G-fold cross validation and so partition the dataset into G subsamples $\mathcal{S}_1, \ldots, \mathcal{S}_G$; we refer to $\mathcal{S}_g$ and $\mathcal{S} - \mathcal{S}_g$ as the gth test and training sets, $g = 1, \ldots, G$. For the SCAD penalty we fixed $\gamma = 3.7$. For a given λ, the *cross-validation statistic* is

$$\widehat{CV}(\lambda) = \sum_{g=1}^{G} \left\{ \log L(\widehat{\theta}_{-g}(\lambda)) - \log L_{-g}(\widehat{\theta}_{-g}(\lambda)) \right\} \ . \tag{2.11}$$

where L_{-g} is the observed data likelihood (2.1) for the gth training dataset and $\widehat{\theta}_{-g}(\lambda)$ is the estimate for the gth training data, obtained through the penalized EM algorithm. The optimal λ maximizes $\widehat{CV}(\lambda)$.

Simulation studies reported in *Appendix 2.B* assess the relative performance of cross-validation, use of the Bayesian information criterion, and the sparse generalized cross-validation (Bradic et al., 2011). While it is difficult to make general statements, the different penalty functions yielded good performance under cross-validation (i.e. good sensitivity for picking up important factors) and small mean squared error (MSE) of the β parameter estimates, with a slightly higher tendency to claim association when there is none). Since there is often strong interest in identifying important variables for further study, it is reasonable to place more importance on the sensitivity and MSE criteria and

so we adopt the standard cross-validation approach to selection of the tuning parameter in the the following empirical studies; this statistic is also used in the R package `glmnet`.

2.3 Design and Interpretation of Simulation Studies

In this section, we report on the results of simulation studies designed to assess the performance of the penalized expectation-maximization algorithm for variable selection with interval-censored data. We consider a sample size of $m = 500$ to correspond roughly to the size of the sample in the psoriatic arthritis study. In the first setting, $p = 100$ and $X_i \sim \text{MVN}_p(0, \Sigma)$ are i.i.d. where the (j, k) element of Σ is $\Sigma_{jk} = \rho^{|j-k|}$, with $\rho = 0.3$ or 0.6 to represent a mild or strong autoregressive dependence respectively, $i = 1, 2, \ldots, m$. The conditional hazard for T_i is based on a Weibull regression model where $h(t|X_i; \theta) = \kappa\eta(\eta t)^{\kappa-1} \exp(X_i'\beta)$. We set $\beta_j = 0.5$ for $j = 1, \ldots, 5$ and $j = 96, \ldots, 100$, so that high values of $X_{i,1}, \ldots, X_{i,5}, X_{i,96}, \ldots, X_{i,100}$ are associated with shorter times to the event, and $\beta_j = 0$, $j = 6, \ldots, 95$ so that $T_i \perp (X_{i,6}, \ldots, X_{i,95})|X_{i,1}, \ldots, X_{i,5}, X_{i,96}, \ldots, X_{i,100}$. The elements of X_i with non-zero coefficients were chosen to give both weak and strong dependence within the set of important covariates.

We consider a study with follow-up planned over $[0, 1]$, where for each of $\kappa = 1.0$ and 1.25, we solve for η so that $P(T_i < 1|X_i = 0; \theta) = 0.95$. We let N_i denote the number of assessments for individual i, generated according to a Poisson distribution with mean μ, truncated to ensure at least one follow-up assessment, given by

$$P(N_i = n|N_i \geq 1; \mu) = \frac{\mu^n \exp(-\mu)}{n! \{1 - \exp(-\mu)\}}, n = 1, \ldots .$$

The n_i inspection times $0 < a_{i1} < \cdots < a_{in_i} < 1$ were then generated uniformly over $[0, 1]$. The left and right endpoints of the censoring interval for individual i are then

$L_i = \max(a_{ij} \cdot I(a_{ij} < t_i))$ and $R_i = \min(a_{ij} \cdot I(a_{ij} > t_i))$ respectively. One hundred datasets were then simulated ($nsim = 100$) for $\mu = 10$ and 20 respectively.

For each dataset, variable selection was carried out based on the penalized expectation-maximization (P-EM) algorithm of Section 2.2 with the LASSO, adaptive LASSO (ALASSO) and SCAD penalty ($\gamma = 3.7$). For each analysis, 5-fold cross validation was carried out to select the unknown tuning parameter. Analyses were conducted based on proportional hazards models with a piecewise constant baseline hazards with four pieces where the break-points were chosen to correspond to the quartiles of the baseline survival function. For comparison with a simple alternative approach, datasets were created by an *ad hoc* mid-point imputation approach (Lindsey and Ryan, 1998) in which event times for individuals with $R_i < \infty$ were taken to be $t_i^* = (L_i + R_i)/2$. The resulting datasets were analysed based on the proportional hazards assumption with piecewise constant baseline hazards with the same break-points as used in the P-EM analyses; the corresponding results are labeled MID. The more traditional methods of variable selection based on forward selection and backward elimination were also considered under the true parametric Weibull regression model where we used $p = 0.10$ for inclusion or removal of terms; the R function `survreg` (R Core Team, 2013; Therneau, 2013) was used in this case as it handles parametric modeling with interval-censored data.

Method		$\mu = 10$ TP(10)	FP(90)	MSE (SD)	$\mu = 20$ TP(10)	FP(90)	MSE (SD)
				Shape parameter: $\kappa = 1$			
LASSO	P-EM	10.00	14.80	0.312 (0.126)	10.00	14.83	0.261 (0.105)
	MID	10.00	13.05	1.346 (0.286)	10.00	12.05	0.912 (0.251)
ALASSO	P-EM	10.00	0.12	0.057 (0.047)	10.00	0.07	0.047 (0.040)
	MID	9.69	0.30	0.953 (0.328)	10.00	1.57	0.499 (0.201)
SCAD	P-EM	9.98	0.36	0.059 (0.073)	9.99	0.24	0.050 (0.048)
	MID	9.39	0.96	0.946 (0.354)	9.91	1.01	0.521 (0.213)
FORWARD		10.00	9.17	0.218 (0.088)	10.00	9.50	0.201 (0.082)
BACKWARD		10.00	15.35	0.322 (0.130)	10.00	14.80	0.289 (0.099)
				Shape parameter: $\kappa = 1.25$			
LASSO	P-EM	10.00	14.88	0.291 (0.118)	10.00	14.13	0.245 (0.109)
	MID	10.00	15.28	1.037 (0.271)	10.00	12.94	0.685 (0.216)
ALASSO	P-EM	9.99	0.23	0.055 (0.050)	10.00	0.08	0.045 (0.031)
	MID	9.75	0.29	0.724 (0.327)	10.00	1.25	0.314 (0.160)
SCAD	P-EM	9.98	0.29	0.055 (0.052)	9.99	0.13	0.044 (0.036)
	MID	9.53	0.76	0.741 (0.336)	9.97	0.91	0.317 (0.167)
FORWARD		10.00	8.66	0.324 (0.089)	10.00	8.81	0.313 (0.089)
BACKWARD		10.00	14.35	0.383 (0.092)	10.00	14.17	0.363 (0.092)

Table 2.1: Empirical results for interval-censored data with normally distributed covariates ($p = 100$, $E(X_{ij}) = 0$, $Var(X_{ij}) = 1$ and $\text{corr}(X_{ij}, X_{ik}) = \rho^{|j-k|}$, where $\rho = 0.5$) summarizing the number of correctly (TP) and incorrectly (FP) selected variables along with the median and the standard deviation (SD) of the mean squared error (MSE); P-EM denotes analyses based on the proposed penalized EM method and MID denotes analyses based on a pseudo-dataset obtained by mid-point imputation; the tuning parameter is selected by five-fold cross-validation.

The number of variables selected was recorded. Among those that are truly associated with the response, the average number selected across all simulated datasets is reported as the mean number of true positive (TP) selections; the correct number of non-zero

coefficients is given in parentheses in the column headings as TP(10). Among the covariates having no (conditional) association with the event time, the number selected for each dataset was averaged and reported as the mean number of false positive (FP) selections; the number of truly independent covariates is given in parentheses as FP(90). These statistics, along with the mean squared error $(\widehat{\beta}-\beta)'\Sigma(\widehat{\beta}-\beta)$, and the empirical standard errors of the mean square error, are reported in Table 2.1 based on 100 simulations.

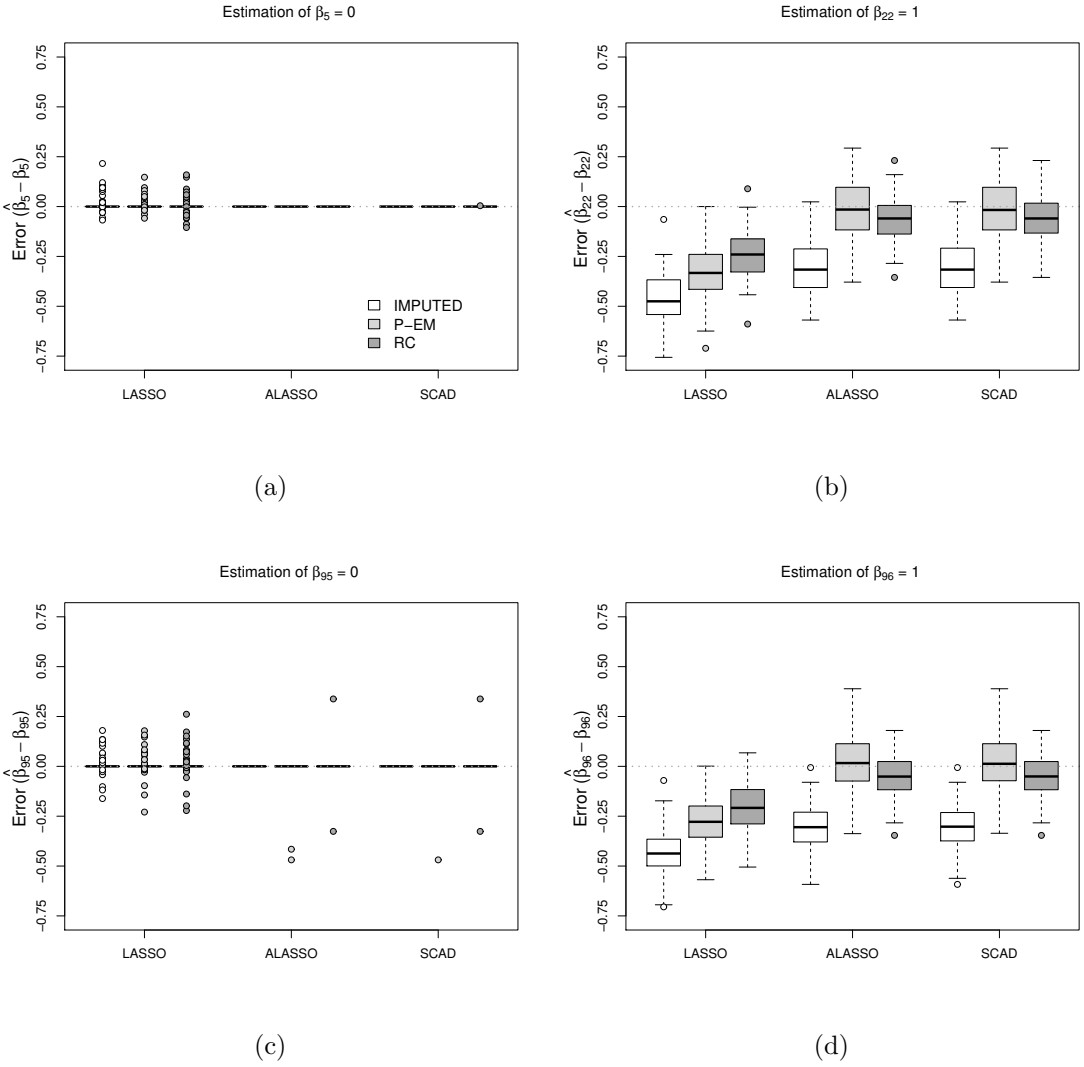
All three penalty functions generally led to selection of the ten covariates associated with the response for the P-EM and MID implementations, with slightly worse performance of the ALASSO and SCAD penalty functions following mid-point imputation. The ALASSO and SCAD penalty functions had the lowest FP values which were lower in the P-EM implementation than following mid-point imputation. For any particular penalty function the mean squared error and the respective standard deviation were always lower when the penalized EM algorithm was used rather than mid-point imputation. These findings point to the advantages of the proposed method which include slightly lower FP values and substantially lower mean squared errors. The forward and backward selection algorithms also featured high FP values. There were little differences between the findings with the exponential ($\kappa = 1$) and Weibull ($\kappa = 1.25$) regression models.

Method		$\mu = 10$ TP(10)	FP(90)	MSE (SD)	$\mu = 20$ TP(10)	FP(90)	MSE (SD)
				Shape parameter: $\kappa = 1$			
LASSO	P-EM	10.00	12.49	0.304 (0.068)	10.00	15.30	0.201 (0.052)
	MID	10.00	17.64	0.690 (0.117)	10.00	19.01	0.436 (0.086)
ALASSO	P-EM	9.88	0.82	0.071 (0.067)	9.98	0.26	0.039 (0.033)
	MID	9.18	0.78	0.491 (0.149)	9.83	0.49	0.255 (0.097)
SCAD	P-EM	9.94	0.54	0.063 (0.063)	10.00	0.10	0.038 (0.031)
	MID	9.02	0.96	0.505 (0.166)	9.79	0.40	0.254 (0.102)
FORWARD		10.00	11.14	0.244 (0.078)	10.00	11.09	0.183 (0.057)
BACKWARD		10.00	15.18	0.299 (0.083)	10.00	14.64	0.231 (0.064)
				Shape parameter: $\kappa = 1.25$			
LASSO	P-EM	10.00	12.04	0.277 (0.064)	10.00	15.65	0.186 (0.053)
	MID	9.99	18.15	0.609 (0.100)	10.00	17.91	0.374 (0.074)
ALASSO	P-EM	9.98	0.59	0.051 (0.042)	10.00	0.22	0.034 (0.023)
	MID	9.59	0.60	0.404 (0.116)	9.97	0.26	0.186 (0.064)
SCAD	P-EM	10.00	0.48	0.053 (0.038)	10.00	0.16	0.033 (0.021)
	MID	9.54	0.93	0.414 (0.118)	9.95	0.42	0.186 (0.064)
FORWARD		10.00	10.86	0.198 (0.060)	10.00	10.81	0.180 (0.045)
BACKWARD		10.00	14.49	0.233 (0.064)	10.00	13.76	0.195 (0.052)

Table 2.2: Empirical results for interval-censored data with correlated binary covariates ($p = 100$, $E(X_{ij}) = 0.2$ and $\text{corr}(X_{ij}, X_{ik}) = \rho^{|j-k|}$ if X_{ij}, X_{ik} are in the same block as discussed in Section 2.3 and $\rho = 0.2$) summarizing the number of correctly (TP) and incorrectly (FP) selected variables along with the median and the standard deviation (SD) of the mean squared error (MSE); P-EM denotes analyses based on the proposed penalized EM method and MID denotes analyses based on a pseudo-dataset obtained by mid-point imputation; the tuning parameter is selected by five-fold cross-validation.

In a second simulation study, we considered correlated binary covariates with $p = 100$ to more closely represent the dimension of the HLA variables in the psoriatic arthritis

study. We set $P(X_{ij} = 1) = 0.20,\ j = 1, \ldots, 100$. For the dependence structure we considered the covariates as arising in ten independent blocks such that the correlation between covariates X_{ij} and X_{ik} within the same block is $\text{corr}(X_{ij}, X_{ik}) = \rho^{|j-k|}$ with $\rho = 0.2$. Ten covariates were identified to have coefficients equal to one, where these were chosen to give a combination of pairwise independence as well as weak, moderate and strong associations between important covariates; all other covariate effects were set to zero. The results displayed in Table 2.2 again demonstrate that all methods tend to select the covariates with the non-zero coefficients on average, although the methods based on the adaptive LASSO and SCAD penalties have negligibly lower TP values. As in the previous simulations, the false positive selection rate is lower with the adaptive LASSO and SCAD penalty functions compared to the LASSO as well as the forward and backward selection algorithms. The respective mean squared errors are always substantially lower in the penalized EM algorithms compared to the respective implementation following mid-point imputation.



(a) (b)

(c) (d)

Figure 2.2: Box plots of the error for the estimated regression coefficients $\widehat{\beta}_k - \beta_k$, $k = 5, 22, 95, 96$, for each penalty function for datasets with correlated binary covariates ($p = 100$) with $\kappa = 1.25$, $\mu = 20$.

Figure 2.2 contains box plots of the errors in estimates (i.e. $\widehat{\beta}_k - \beta_k$) for four of the

hundred coefficients in the setting with binary covariates, $\kappa = 1.25$, and $\mu = 20$; β_5 and β_{95} (both zero) and β_{22} and β_{96} (both 1.0). For each penalty function the estimates for the P-EM and MID methods are displayed, along with estimates from an analysis using the true failure time subject only to administrative right censoring (RC) at $C = 1$; the latter analysis is only possible in a simulation study, but is presented for comparison purposes since it provides a natural benchmark for assessing the performance of the proposed algorithm for interval-censored data. It is important to note that different datasets are used for the P-EM, MID and RC analyses, with only the former corresponding to the observed data.

In *Appendix 2.A*, we present the results of further simulation studies with multivariate normal and correlated binary covariates when $p = 10$. Here we consider analyses with an exponential (time homogeneous) regression model and a piecewise constant baseline hazards (4 pieces) model. The former is included to examine the effect of having a more elaborate (four piece) baseline hazard when a single piece is sufficient as is the case when $\kappa = 1.0$, as well as the effect of gross misspecification of the baseline hazard when $\kappa = 1.25$. When $\kappa = 1.0$ and the P-EM algorithm is used, the PWC-4 model yields a very slightly higher MSE than was seen for the exponential model, but the results suggest there is little price to pay when the piecewise constant model is used unnecessarily.

When $\kappa = 1.25$, the piecewise constant model (PWC-4) had a slightly lower rate of false positive selections and a lower MSE than the exponential model. A similar study was conducted with binary covariates ($p = 10$) with findings that suggest that the adaptive LASSO and SCAD penalties are again preferable to the LASSO since they generally lead to smaller MSE; among these two methods the relative performance tends to depend on the criteria used (TP, FP or MSE) but they appear broadly comparable overall.

2.4 HLA Markers and Risk of Arthritis Mutilans

Patients are classified as suffering from arthritis mutilans upon the occurrence of their fifth damaged joint, and interest lies in identifying which among the 76 human leukocyte antigen markers are associated with increased risk of reaching this stage from the time of diagnosis with psoriatic arthritis. The first, second and third quartiles for the closed censoring intervals for the 97 individuals known to have developed arthritis mutilans were 2.50, 8.06 and 15.00 years respectively. These quantiles are much wider than one might expect from a protocol in which radiological assessments are to be scheduled every two years because of the variation between individuals in the propensity to attend the clinic, as well as the potentially long delay from the onset of psoriatic arthritis to clinic entry; see Figure 2.1a. We also remark that the proportion of individuals generating interval-censored times to arthritis mutilans is smaller than that represented in the simulation study, and that the variability in the width of the censoring intervals is considerable; the algorithm can accommodate this setting.

We seek to identify which of the 76 human leukocyte antigens have prognostic value, while controlling for 6 clinical predictors including age at clinic entry, sex, age at onset of psoriasis, age at onset of PsA, family history of psoriasis (yes/no), and family history of psoriatic arthritis (yes/no). We report here on the results of applying the penalized EM algorithm to the interval-censored time of arthritis mutilans among patients in the University of Toronto Psoriatic Arthritis Clinic, using the LASSO, adaptive LASSO and SCAD penalty functions. For comparison purposes, results are also reported based on a right-censored dataset obtained by using midpoint imputation (MID) as examined in the simulation studies. Given the findings from the simulation studies; however, we restrict our attention primarily to the results from the penalized expectation-maximization proce-

dure. The standard errors of the estimates are calculated using the bootstrap (Efron and Tibshirani, 1994); details are given in *Appendix 2.C*.

The break-points for the piecewise constant hazard functions were chosen based on the nonparametric estimate of the marginal cumulative probability distribution function for the time from disease onset to arthritis mutilans; see Figure 2.1b. The cumulative probability is about 35% over 28 years so the break-points chosen were 6.5, 10.5, 18, and 22 years corresponding to the cumulative probabilities of 7%, 14%, 21% and 28%.

The union of all HLA variables selected by any method are listed in Table 2.3, where it can be seen that the SCAD penalty function with the P-EM procedure selected the fewest HLA markers including HLA-A11, HLA-A29, HLA-B27 and HLA-DQB1-02; HLA-B27 and HLA-DQB1-02 are two factors well known to incur increased risk of joint damage and we found that the presence of HLA-A11 and HLA-A29 has a protective effect. Under the P-EM algorithm the LASSO penalty function also selected HLA-C04, and the corresponding implementation of the ALASSO further selected HLA-A25, HLA-A30 and HLA-DRB1-10. With the ALASSO penalty the same variables were selected whether the P-EM or mid-point imputation was used; for the other penalty functions more variables were selected under mid-point imputation than with the P-EM procedure, as found in the empirical investigations. The findings are in broad agreement with those from recent analyses (Chandran et al., 2012) and a validation exercise is currently underway involving three independent cohorts from Spain, Ireland and Newfoundland, Canada. The empirical correlations among the union set of all variables selected by any method range from -0.105 to 0.198.

	LASSO				ALASSO				SCAD			
	P-EM		MID		P-EM		MID		P-EM		MID	
HLA Marker	β	s.e.(β)	β	s.e.(β)	β	s.e.(β)	β	s.e.(β)	β	s.e.(β)	β	s.e.(β)
HLA-A11	-0.135	0.199	-0.280	0.263	-0.516	0.629	-0.556	0.836	-1.021	0.746	-0.922	0.947
HLA-A25			-0.232	0.288	-3.265	0.707	-3.229	1.529				
HLA-A29	-0.216	0.254	-0.502	0.353	-1.388	1.284	-1.385	1.440	-1.605	2.376	-1.658	2.482
HLA-A30			0.101	0.260	0.494	0.417	0.494	0.525				
HLA-B27	0.249	0.232	0.397	0.272	0.588	0.356	0.595	0.547	0.763	0.312	0.725	0.425
HLA-C04	-0.012	0.134	-0.170	0.233	-0.578	0.492	-0.569	1.086			-0.637	0.611
HLA-DQB1-02	0.134	0.164	0.270	0.205	0.514	0.307	0.503	0.540	0.609	0.276	0.623	0.415
HLA-DRB1-10					-2.713	1.007	-2.714	1.725				

Table 2.3: Selected HLA markers and their effects obtained by variable selection with interval-censored data disease progression data in psoriatic arthritis using the LASSO, ALASSO and SCAD penalty functions.

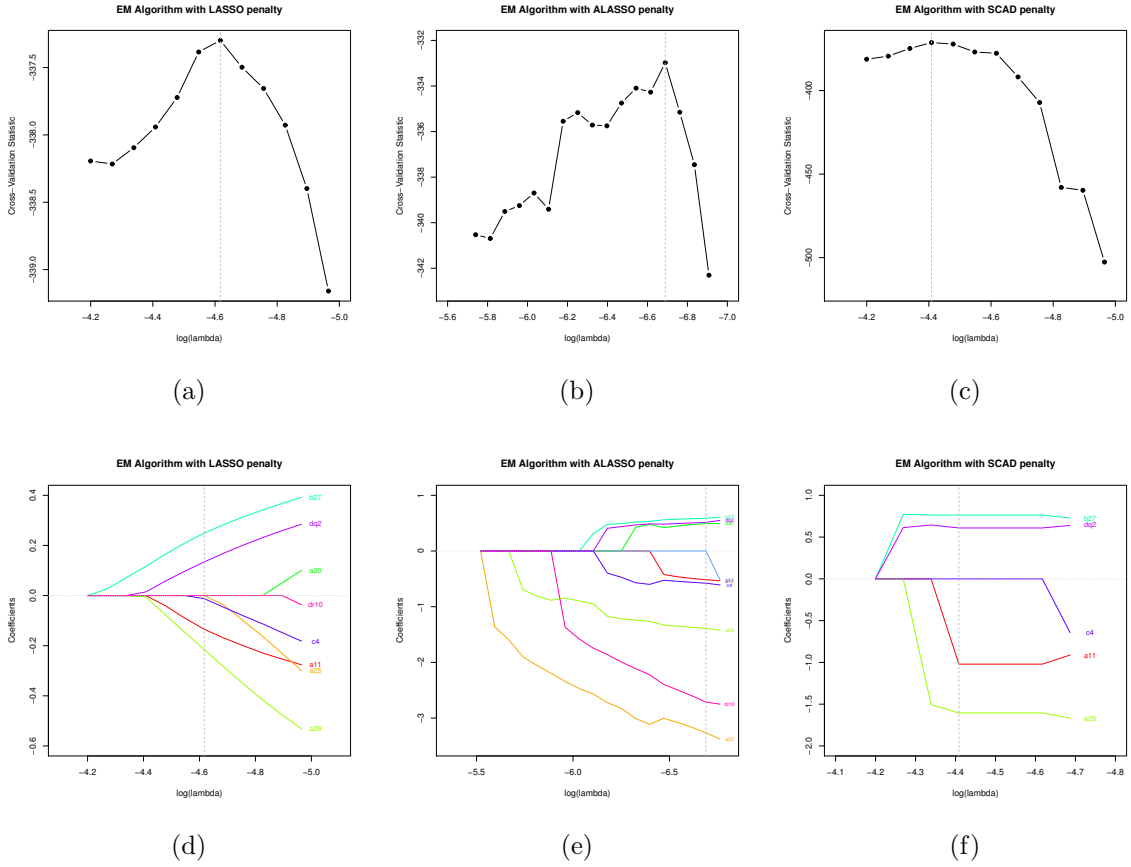


Figure 2.3: Plots of the cross validation statistics and shrinkage of coefficients in the PsA dataset based on piecewise constant hazard model via EM algorithm with the LASSO, ALASSO and SCAD penalty functions.

The upper panels of Figure 2.3 contain plots of the cross-validation statistic to reveal how the optimal values of the tuning parameters are found for the LASSO, adaptive ALASSO and SCAD functions; the plots in the lower panels of Figure 2.3 give the profile plots of the coefficients, showing the degree of shrinkage and selection of covariates as a function of the tuning parameter. The stage at which each variable is selected conveys

the relative importance of the covariates; the optimal value of the tuning parameter is designated by the vertical dotted lines.

2.5 Summary of Findings on Penalized Regression for Interval-Censored Data

Thus far in this chapter we have proposed a simple adaptation of existing algorithms for variable selection to deal with interval-censored failure time data. A complete data log-likelihood form based on a proportional hazards model with a piecewise constant baseline hazard is augmented by including one of several possible penalty terms. The simulation studies showed that the proposed algorithm led to better performance for each penalty function compared to simple methods using mid-point imputation. We experienced no convergence problems with the penalized expectation-maximization algorithm; Wu (1983) should help assess whether this can be relied upon generally. The adaptive LASSO, as implemented here with iteratively updated weights, had perhaps the best performance. The relative performance of the different penalty functions depended heavily on the method for selecting the optimal tuning parameter in the penalty functions. It can be seen in Table B.2 of the *Appendix 2.B*, for example, that the performance of the LASSO in terms of FP was much better when tuning parameter λ was chosen by BIC or SGCV. The purpose of this chapter is not to carry out an exhaustive study of variable selection techniques based on the different penalty functions, but further study of the various options for choosing the tuning parameters seems worthwhile.

An application of PsA data was conducted, and the results agree quite well with the previous analysis. Lockhart et al. (2014) point out the properties of coefficients obtained

following variable selection are not well understood. In *Appendix 2.C* we explore techniques for variance estimation following variable selection, but we rely on bootstrap standard errors in the application.

The piecewise exponential model is a simple, flexible and weakly parametric approach to dealing with interval-censored data. We set $K = 4$, following the observation of Lawless and Zhan (1998) that a modest number of pieces is usually sufficient, particularly when inferences about covariate effects are of greatest interest. More flexible semiparametric methods could be considered in this setting, including methods based on local likelihood (Betensky et al. 2002; Braun et al. 2005) or penalized splines (Cai and Betensky, 2003). These, and other semiparametric methods (Sun, 2006), may offer a more suitable framework for studying the limiting behaviour of these algorithms and the resultant estimators.

A natural extension of this work is for the analysis of recurrent events observed subject to interval-censoring. In many clinical settings, events can recur over time. In osteoporosis, for example, patients can have fractures repeatedly over time and these may only be detectable upon periodic radiographic examination. In the psoriatic arthritis clinic, when interest lies in modeling the cumulative number of damaged joints, this count is often based upon damage scores determined by radiographic examination. The resulting data, consisting of a series of assessment times and counts representing the number of events occurring between consecutive assessments, is often called panel count data (Sun and Kalbfleisch, 1995). Lawless and Zhan (1998) develop the likelihood and estimating functions for the analysis of such data for mixed Poisson models with piecewise-constant rate functions (Cook and Lawless, 2007). The former can be naturally adopted to allow variable selection based on penalized likelihood for recurrent event data. Given the individual patient level random effect, the penalized likelihood has a similar form to the one in Section 2.2. While the observed data likelihood can be penalized, a complete data likelihood involving a more

detailed recording of the counts and the patient level random effect is very appealing and can exploit existing software. See He et al. (2009) for a semiparametric implementation of a similar algorithm. Tong et al. (2009) develop penalized estimating functions for variable selection with panel count data and Wu and He (2012) propose and study a fast and efficient coordinate ascent algorithm for the same problem.

2.6 Penalized Regression for Truncated and Censored Data

2.6.1 The Motivating Study and Sample Selection Conditions

The goal of this research is to identify genetic factors associated with rapid onset of psoriatic arthritis (PsA) among individuals with psoriasis (Ps) using data from disease registries with different selection conditions: the University of Toronto Psoriasis Clinic (UTPC) and the University of Toronto Psoriatic Arthritis Clinic (UTPAC).

Figure 2.4 contains two Lexis diagrams characterizing the selection criteria for patients into the UTPC and UTPAC cohorts for a hypothetical individual; the horizontal axis represents the timing of events in calendar time while the vertical axis conveys the times since the development of psoriasis. We let B_i denote the date of birth of individual i, E_{i0} denote the calendar time of the onset of psoriasis, and let E_{i1} denote the calendar time psoriatic arthritis developed. The time from the onset of Ps to the onset of PsA is denoted $T_i = E_{i1} - E_{i0}$.

The calendar time at which individuals are screened is denoted by A_0. For the UTPC cohort individuals are required to have psoriasis at the time of screening but cannot have

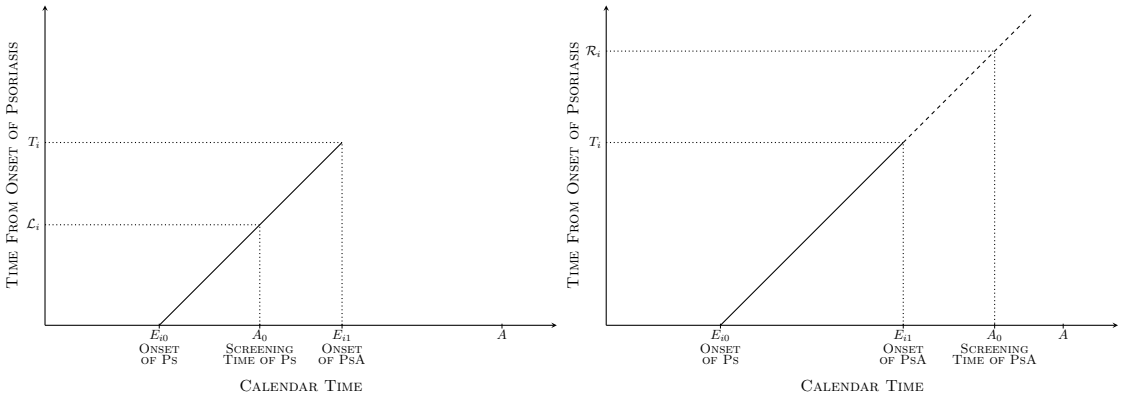


Figure 2.4: Lexis diagrams of the calendar times of birth (B), onset of psoriasis (E_0) and onset of psoriatic arthritis (E_1), along with screening times (A_0) for UTPC (left panel) and the UTPAC (right panel).

developed PsA, so patients are recruited to this registry subject to the constraint $E_{i0} < A_0 < E_{i1}$ (left panel Figure 2.4). Given E_{i0}, this can be equivalently expressed as the constraint $T_i \geq \mathcal{L}_i$ where $\mathcal{L}_i = A_0 - E_{i0}$ is the left-truncation time for T_i. For the PsA cohort, only screened subjects who are determined to have PsA are included in the registry, so in this cohort, subjects are sampled subject to the constraint $E_{i1} < A_0$, or equivalently given E_{i0} subject to $T_i \leq \mathcal{R}_i$ where $\mathcal{R}_i = A_0 - E_{i0}$ is the right-truncation time for T_i (right panel Figure 2.4). To unify the notation for the two cohorts we let $\mathcal{A}_i = [\mathcal{L}_i, \mathcal{R}_i)$ denote the truncation interval for individual i, such that $0 < \mathcal{L}_i < \mathcal{R}_i = \infty$ for individuals in the UTPC, and $0 = \mathcal{L}_i < T_i < \mathcal{R}_i$ for individuals in the UTPAC.

Upon recruitment to each cohort patients are examined intermittently and we let $A_{i1} < A_{i2} < \cdots < A_{in_i}$ denote the calendar times of n_i follow-up assessments for individual i realized over $[A_0, A]$ where A is the date the databases are locked for analysis. If $E_{i1} \in [A_{i,j-1}, A_{ij}]$ for some $j = 1, \ldots, n_i$ then PsA is known to have developed, but it is subject to interval-censoring. We let $\mathcal{C}_i = [L_i, R_i)$ denote the interval containing T_i where $L_i =$

$A_{i,j-1} - E_{i0}$ and $R_i = A_{ij} - E_{i0}$. When T_i is interval-censored $0 < L_i < R_i < \infty$, if it is right-censored $R_i = \infty$, and if T_i is observed then $L_i = R_i = T_i$. We take the dates of diagnosis of psoriatic arthritis in medical records as known; with respect to the onset time of PsA only the retrospective data are used from the UTPAC. If $X_i = (X_{i1}, \ldots, X_{ip})'$ denotes a $p \times 1$ covariate vector associated with individual i, the observed data from individual i are denoted by $D_i = (\mathcal{A}_i, \mathcal{C}_i, X_i)$ and the observed data for a pooled sample of size m is $D = \{D_i, i = 1, \ldots, m\}$.

Under the proportional hazards model $h(t|X_i;\theta) = h_0(t;\alpha)\exp(X_i'\beta)$, α parameterizes the baseline hazard, $\beta = (\beta_1, \ldots, \beta_p)'$, and we let $\theta = (\alpha', \beta')'$. With independent truncation and independent, non-informative censoring (Klein and Moeschberger, 2003), the partial likelihood given X_i is

$$L(\theta) \propto \prod_{i=1}^{m} \frac{\mathcal{F}(L_i|X_i;\theta) - \mathcal{F}(R_i|X_i;\theta)}{\mathcal{F}(\mathcal{L}_i|X_i;\theta) - \mathcal{F}(\mathcal{R}_i|X_i;\theta)} ,$$

where $H(t|X_i;\theta) = \int_0^t h(s|X_i;\theta)ds$ and $\mathcal{F}(t|X_i;\theta) = \exp\{-H(t|X_i;\theta)\}$ is the survivor function.

When the dimension p is large it is customary to adopt some form of penalty for model complexity to help in the selection of important variables for further investigation. Most such penalized log-likelihoods can be written in the form $\ell_{\text{PEN}}(\theta) = m^{-1}\log L(\theta) - p_{\gamma,\lambda}(\beta)$, where $p_{\gamma,\lambda}(\beta)$ determines the extent of the penalty for each value of β, modulated by the tuning parameters (γ, λ). We consider three penalty functions, the LASSO, the adaptive LASSO and the SCAD penalties as described in the previous sections.

2.6.2 A Turnbull-Type EM Algorithm

Let J_i denote the number of "missing" individuals who have the same characteristics as the ith sampled individual except they did not satisfy the selection criteria (i.e. their event times fall in $\mathcal{A}_i^c$). We further let $T_{ij} \in \mathcal{A}_i^c$ be the event time of the jth unselected individual corresponding to individual i, so a Turnbull-type (Turnbull, 1976) complete data likelihood is

$$L_C(\theta) \propto \prod_{i=1}^{m} \left\{ h(T_i|X_i;\theta) \exp(-H(T_i|X_i;\theta)) \prod_{j=1}^{J_i} h(T_{ij}|X_i;\theta) \exp(-H(T_{ij}|X_i;\theta)) \right\} .$$

The reason for considering this form is that by introducing the unobserved failure times and adopting a weakly parametric piecewise constant baseline hazard model via EM algorithm, the maximization step of the complete data likelihood will be simplified.

Under a weakly parametric piecewise constant baseline hazard function, the number and location of break-points at which the baseline hazard changes value must be specified. If $0 = b_0 < b_1 < \cdots < b_{K-1} < b_K = \infty$ denote K break-points, we let $h_0(s;\alpha) = \exp(\alpha_k)$, for $s \in \mathcal{B}_k = [b_{k-1}, b_k)$, $k = 1, \ldots, K$. Let $D_k(u) = I(u \in \mathcal{B}_k)$ denote whether or not the time u is in the interval $\mathcal{B}_k$ and $W_k(u) = \int_0^u D_k(s)ds$ denote the duration of $[0, u)$ over interval k, $k = 1, \ldots, K$. Then under the piecewise constant model and given a covariate vector X_i, the complete data log-likelihood would be

$$\begin{aligned} \log L_C(\theta) \propto \sum_{i=1}^{m} \sum_{k=1}^{K} \Bigg\{ & D_k(T_i) \left(\alpha_k + X_i'\beta\right) - W_k(T_i) \exp(\alpha_k + X_i'\beta) \\ & + \sum_{j=1}^{J_i} \left[D_k(T_{ij}) \left(\alpha_k + X_i'\beta\right) - W_k(T_{ij}) \exp(\alpha_k + X_i'\beta) \right] \Bigg\} . \end{aligned} \tag{2.12}$$

If $Z_{ik\ell} = I(k = \ell)$ and $Z_{ik} = (Z_{ik1}, \ldots, Z_{ikK})'$ denotes the corresponding vector of indicator functions, $k = 1, \ldots, K$; thus $Z_{i1} = (1, 0, \ldots, 0)'$, $Z_{i2} = (0, 1, \ldots, 0)'$, ...,

$Z_{ik} = (0, 0, \ldots, 1)'$. Then if $\alpha = (\alpha_1, \ldots, \alpha_K)'$ and $\theta = (\alpha', \beta')'$, we can write

$$\log L_C(\theta) = \sum_{i=1}^{m} \log L_{Ci}(\theta) \ ,$$

where upon letting $\bar{X}_{ik} = (Z'_{ik}, X'_i)'$ we can write $\log L_{Ci}(\theta)$ as

$$\sum_{k=1}^{K} \left\{ D_k(T_i)\bar{X}'_{ik}\theta - W_k(T_i)\exp(\bar{X}'_{ik}\theta) + \sum_{j=1}^{J_i} \left[D_k(T_{ij})\bar{X}'_{ik}\theta - W_k(T_{ij})\exp(\bar{X}'_{ik}\theta) \right] \right\} \ .$$

At the E-step of the EM algorithm, the conditional expectation of the penalized complete data log-likelihood function at the $(r+1)$st iteration is evaluated as

$$Q_{\text{PEN}}(\theta; \theta^{(r)}) = \sum_{i=1}^{m} Q_i(\theta; \theta^{(r)}) - p_{\gamma,\lambda}(\beta) \ , \tag{2.13}$$

where $Q_i(\theta; \theta^{(r)}) = E\left\{ \log L_{Ci}(\theta) | D; \theta^{(r)} \right\}$ and $\theta^{(r)}$ is estimated by maximizing $Q_{\text{PEN}}(\theta; \theta^{(r-1)})$. The required conditional expectations are therefore $\widehat{\Delta}_{ik}^{(r)} = E[D_k(T_i)|D_i; \theta^{(r)}]$, $\widehat{\mathcal{S}}_{ik}^{(r)} = E[W_k(T_i)|D_i; \theta^{(r)}]$, $\widehat{\iota}_{ik}^{(r)} = E[D_k(T_{ij})|D_i; \theta^{(r)}]$, $\widehat{\omega}_{ik}^{(r)} = E[W_k(T_{ij})|D_i; \theta^{(r)}]$ and $\mathcal{J}_i^{(r)} = E[J_i|D_i; \theta^{(r)}]$.

Let $\mathcal{C}_{ik} = \mathcal{C}_i \cap \mathcal{B}_k = [L_{ik}, R_{ik})$ denote the sub-interval of the censoring interval $\mathcal{C}_i$ contained within $\mathcal{B}_k$. When $\mathcal{C}_{ik} = \emptyset$, the required expectations are relatively easy to compute since, for instance, it is clear that $D_k(t_i) = 0$ and $\widehat{\Delta}_{ik}^{(r)} = 0$. Moreover, if $b_k < L_i$, then it is known that individual i was at risk for the entire interval $\mathcal{B}_k$ so $W_k(t_i) = \widehat{\mathcal{S}}_{ik}^{(r)} = b_k - b_{k-1}$, and if $R_i < b_{k-1}$, then $W_k(t_i) = \widehat{\mathcal{S}}_{ik}^{(r)} = 0$ since they are known to have failed prior to the start of interval $\mathcal{B}_k$. If $\mathcal{C}_{ik} \neq \emptyset$,

$$\widehat{\Delta}_{ik}^{(r)} = \frac{\mathcal{F}(L_{ik}|X_i; \theta^{(r)}) - \mathcal{F}(R_{ik}|X_i; \theta^{(r)})}{\mathcal{F}(L_i|X_i; \theta^{(r)}) - \mathcal{F}(R_i|X_i; \theta^{(r)})} \ , \tag{2.14}$$

$$\widehat{\mathcal{S}}_{ik}^{(r)} = \max(L_i - b_{k-1}, 0) + \int_{L_{ik}}^{R_{ik}} \frac{\mathcal{F}(s|X_i; \theta^{(r)})}{\mathcal{F}(L_i|X_i; \theta^{(r)}) - \mathcal{F}(R_i|X_i; \theta^{(r)})} \mathrm{d}s \ , \tag{2.15}$$

where $\mathcal{F}(t|X_i; \theta) = \exp\left\{ -\left(\sum_{k=1}^{K} \exp(\alpha_k) W_k(t) \right) \exp(X'_i\beta) \right\}$.

Let $\mathcal{A}_{ik} = \mathcal{A}_i^c \cap \mathcal{B}_k = [\mathcal{L}_{ik}, \mathcal{R}_{ik})$ be the sub-interval of the complement of the truncation interval $\mathcal{A}_i^c$ contained within $\mathcal{B}_k$, if $\mathcal{A}_{ik} = \emptyset$, then $\widehat{\iota}_{ik}^{(r)} = 0$. Moreover, if $\mathcal{A}_i^c = [0, \mathcal{L}_i)$, then $\widehat{\omega}_{ik}^{(r)} = 0$ since they are known to have failed prior to the start of interval, and if $\mathcal{A}_i^c = (\mathcal{R}_i, \infty)$, then the individual i was at risk for the entire interval $\mathcal{B}_k$ so $\widehat{\omega}_{ik}^{(r)} = b_k - b_{k-1}$. If $\mathcal{A}_{ik} \neq \emptyset$,

$$\widehat{\iota}_{ik}^{(r)} = \frac{\mathcal{F}(\mathcal{L}_{ik}|X_i;\theta^{(r)}) - \mathcal{F}(\mathcal{R}_{ik}|X_i;\theta^{(r)})}{1 - \mathcal{F}(\mathcal{L}_i|X_i;\theta^{(r)}) + \mathcal{F}(\mathcal{R}_i|X_i;\theta^{(r)})} , \tag{2.16}$$

$$\widehat{\omega}_{ik}^{(r)} = \mathcal{L}_{ik} - b_{k-1} + \int_{\mathcal{L}_{ik}}^{\mathcal{R}_{ik}} \frac{\mathcal{F}(s|X_i;\theta^{(r)})}{1 - \mathcal{F}(\mathcal{L}_i|X_i;\theta^{(r)}) + \mathcal{F}(\mathcal{R}_i|X_i;\theta^{(r)})} \mathrm{d}s . \tag{2.17}$$

Also

$$\widehat{\mathcal{J}}_i^{(r)} = E[J_i|D_i;\theta^{(r)}] = \frac{1 - \mathcal{F}(\mathcal{L}_i|X_i;\theta^{(r)}) + \mathcal{F}(\mathcal{R}_i|X_i;\theta^{(r)})}{\mathcal{F}(\mathcal{L}_i|X_i;\theta^{(r)}) - \mathcal{F}(\mathcal{R}_i|X_i;\theta^{(r)})} . \tag{2.18}$$

Given these results, (2.7) can be written more explicitly as

$$\sum_{i=1}^{m}\sum_{k=1}^{K}\left\{ \left[\widehat{\Delta}_{ik}^{(r)}\bar{X}'_{ik}\theta - \widehat{\mathcal{S}}_{ik}^{(r)}\exp(\bar{X}'_{ik}\theta)\right] + \widehat{\mathcal{J}}_i^{(r)}\left[\widehat{\iota}_{ik}^{(r)}\bar{X}'_{ik}\theta - \widehat{\omega}_{ik}^{(r)}\exp(\bar{X}'_{ik}\theta)\right]\right\} - p_{\gamma,\lambda}(\beta) . \tag{2.19}$$

Since (2.19) has the form of a penalized Poisson likelihood, the M-step can be carried out using software for penalized Poisson regression. This can be implemented by creating an augmented pseudo-dataset with individual i contributing up to K lines with weight 1 and K lines (for the corresponding unselected individuals) with weight $\widehat{\mathcal{J}}_i^{(r)}$, $i = 1, \ldots, m$.

Classical variable selection methods are often based on the Akaike information criterion (AIC) or the Bayesian information criterion (BIC), while more recently cross-validation (CV) and generalized cross-validation (GCV) techniques have been advocated. The traditional G-fold CV statistic is defined as $\widehat{CV}(\lambda) = \sum_{g=1}^{G}[\log L(\widehat{\theta}_{-g}(\lambda)) - \log L_{-g}(\widehat{\theta}_{-g}(\lambda))]$ where L_{-g} is the observed data likelihood for the gth training dataset and $\widehat{\theta}_{-g}(\lambda)$ is the

estimate for the gth training data, obtained through the EM algorithm; the optimal λ maximizes $\widehat{CV}(\lambda)$.

2.6.3 Simulation Studies and Application to the Psoriasis and Psoriatic Arthritis Registries

We considered a sample size of $m = 1200$ with $m_1 = 400$ of the subjects left-truncated and $m_2 = 800$ right-truncated and the number of covariates is $p = 100$. X_{ij}'s are binary covariates with $P(X_{ij} = 1) = 0.5$, $i = 1, \ldots, m, j = 1, \ldots, p$. There are eight covariates specified to have coefficients not equal to zero and all other covariate effects were set to zero, that is $\beta_j = \log(2) = 0.6931$, $j = 1, 2, 9, 10$ and $\beta_j = \log(0.5) = -0.6931$, $j = 17, 18, 19, 20$ and $\beta_j = 0$, otherwise. The conditional hazard for T_i is based on a Weibull regression model where

$$h(t|X_i; \theta) = \kappa\eta(\eta t)^{\kappa-1} \exp(X_i'\beta) \ ,$$

where $\kappa = 1.25$. We consider a study with median event time equal to 1, thus for each of $\kappa = 1$ and 1.25, we solve for η so that

$$P(T_i < 1; \theta) = E_X \left[P(T_i < 1|X; \theta)\right] = 0.5 \ .$$

Let t_{Q25}, t_{Q50} and t_{Q75} be the quartiles of the marginal distribution of T_i and the truncation times are drawn from these quartiles with equal probabilities. For each subject i, it has either a left-truncated right-censored event time (Ps cohort) or a right-truncated event time (PsA cohort).

For the ith subject, $i = 1, \ldots, m_1$, which are subject to left-truncation, we generate the left-truncation time $\mathcal{L}_i$ which is randomly drawn from the quartiles with equal probabilities. To ensure the sample covariate distribution is compatible with the truncation

	AIC			*BIC*			*CV*		
Method	TP(8)	FP(92)	MSE (SD)	TP(8)	FP(92)	MSE (SD)	TP(8)	FP(92)	MSE (SD)
				Shape parameter $\kappa = 1.25$					
				Interval-Censored Left-Truncated and Right-Truncated Data					
LASSO	8.00	15.95	0.120 (0.042)	7.94	3.02	0.224 (0.087)	8.00	16.37	0.120 (0.041)
ALASSO	8.00	10.40	0.146 (0.051)	8.00	0.52	0.034 (0.022)	7.98	0.41	0.029 (0.032)
SCAD	8.00	10.44	0.149 (0.051)	8.00	0.54	0.040 (0.021)	7.98	0.34	0.029 (0.030)
				Right-Censored Left-Truncated and Right-Truncated Data					
LASSO	8.00	17.60	0.113 (0.039)	7.98	2.94	0.224 (0.078)	8.00	16.06	0.119 (0.039)
ALASSO	8.00	10.86	0.152 (0.054)	8.00	0.51	0.035 (0.021)	7.96	0.27	0.026 (0.031)
SCAD	8.00	10.61	0.155 (0.052)	8.00	0.50	0.036 (0.020)	7.98	0.26	0.028 (0.025)
				Shape parameter $\kappa = 1$					
				Interval-Censored Left-Truncated and Right-Truncated Data					
LASSO	8.00	19.69	0.098 (0.030)	8.00	3.27	0.181 (0.065)	8.00	17.09	0.105 (0.032)
ALASSO	8.00	11.42	0.165 (0.063)	8.00	0.61	0.030 (0.028)	8.00	0.47	0.023 (0.032)
SCAD	8.00	11.37	0.166 (0.062)	8.00	0.65	0.032 (0.030)	7.97	0.53	0.023 (0.042)
				Right-Censored Left-Truncated and Right-Truncated Data					
LASSO	8.00	19.21	0.098 (0.030)	8.00	3.38	0.180 (0.070)	8.00	17.09	0.102 (0.033)
ALASSO	8.00	11.49	0.163 (0.063)	8.00	0.64	0.031 (0.029)	7.97	0.48	0.023 (0.042)
SCAD	8.00	11.42	0.160 (0.063)	8.00	0.63	0.029 (0.029)	7.99	0.46	0.024 (0.036)

Table 2.4: Empirical results for dataset (33.33% left-truncation and 66.67% right-truncation) with binary covariates ($p = 100$, $P(X_{ij} = 1) = 0.5$) summarizing the number of correctly (TP) and incorrectly (FP) selected variables along with the median and the standard deviation (SD) of the mean squared error (MSE). The tuning parameter is selected by AIC, BIC or CV. Sample size $m = 1200$, and $nsim = 100$ simulations.

scheme, we generate X_i using the conditional distribution $P(X_i|T_i > \mathcal{L}_i)$. We then generate $U_i \sim \text{Uniform}(0,1)$ and solve for the event time T_i that satisfies $P(T \geq T_i|T \geq \mathcal{L}_i, \mathcal{L}_i = l_i, X_i; \theta) = U_i$. For the ith subject, $i = m_1 + 1, \ldots, m$, whose times are subject to right-truncation, we generate the right-truncation time $\mathcal{R}_i$ uniformly from the quartiles, X_i is generated from $P(X_i|T_i < \mathcal{R}_i)$, solve for T_i in the constraint $P(T < T_i|T < \mathcal{R}_i, \mathcal{R}_i = r_i, X_i; \theta) = U_i$ where $U_i \sim \text{Uniform}(0,1)$. We consider this study with duration of follow-up planned to be $\tau = A - A_0$, where τ is obtained from $P(T \geq \mathcal{L}_i + \tau|T \geq \mathcal{L}_i; \theta) = 0.5$. For simplicity, we consider a fixed number of inspections $n_i = 5$, $i = 1, \ldots, m$, and the follow-up inspection times are generated uniformly from $[\mathcal{L}_i, \mathcal{L}_i + \tau]$, $j = 1, \ldots, 5$, $i = 1, \ldots, m$.

For each dataset, variable selection was carried out based on the penalized EM (P-EM) algorithm with the LASSO, adaptive LASSO (ALASSO) and SCAD ($\gamma = 3.7$) penalty functions. The tuning parameter was selected in each case using the AIC, the BIC or using a 5-fold cross-validation statistic. Analyses were conducted based on proportional hazards models with a piecewise constant baseline hazards; hazard functions with four pieces (PWC-4) where the break-points were based on the quantiles of the baseline survival function.

Table 2.4 displays the performance of LASSO, ALASSO and SCAD for each method of selecting the tuning parameter in the setting with some trend in the baseline hazard and for a time homogeneous model. The probability that an important variable is appropriately selected is generally very high for all methods, but false positive rates are quite high under the LASSO penalty regardless of how the tuning parameter is selected; all methods have high false positive rates when AIC is used for the selection of the tuning parameter. The ALASSO and SCAD penalty functions perform very well when the tuning parameter is selected by BIC or 5-fold cross-validation; the performance is slightly better for the CV than with the BIC criterion.

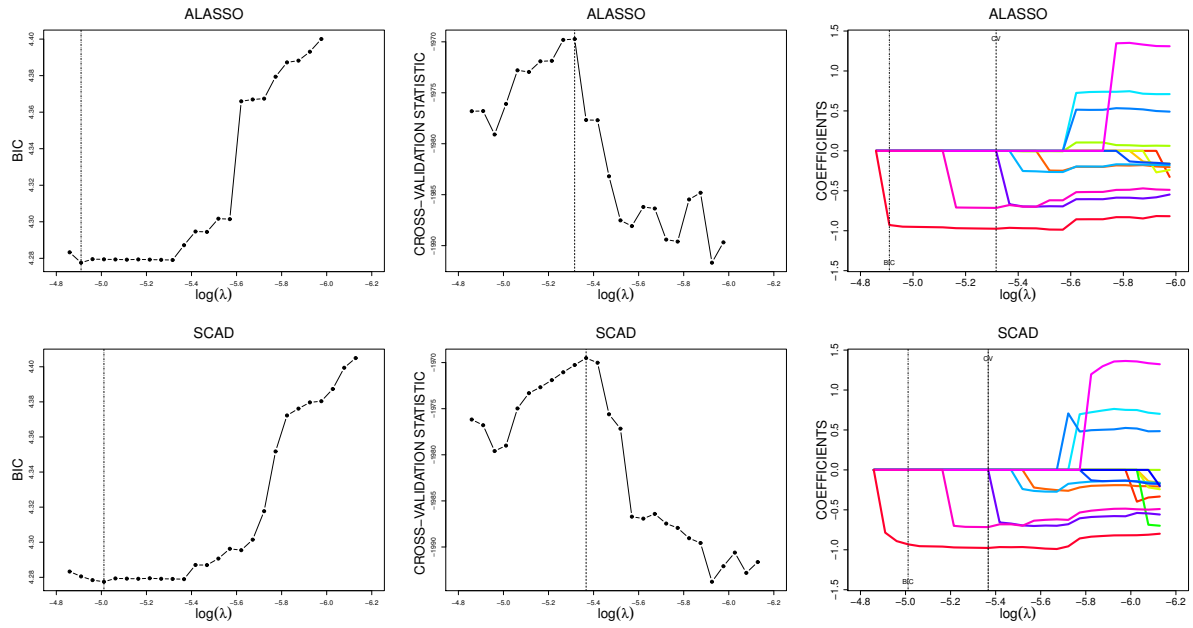


Figure 2.5: Plots of the BIC (1st column), 5-fold cross-validation statistic (2nd column) and shrinkage estimates of coefficients (3rd column) against the tuning parameter from penalized regression of the PsA dataset based on a piecewise constant hazard model (PWC-4) fitted via an EM algorithm with the LASSO, ALASSO or SCAD penalty. The fixed covariates are gender and the onset age of psoriasis.

The data from the UTPC and UTPAC are comprised of 338 and 603 individuals with left- and right-truncated PsA onset times respectively along with data on 76 human leukocyte antigen (HLA) markers. Among the 338 individuals in the UTPC cohort 38 yielded onset dates for psoriatic arthritis. Given the high false positive selection rate of the LASSO and of all methods when the tuning parameter is selected based on the AIC criterion, in this application we use the ALASSO and SCAD penalty functions and select the tuning parameter based on the BIC and 5-fold CV statistic. The basic model involves a piecewise (4-piece) constant baseline hazard and all models control for age and gender.

The findings based on the BIC suggest HLA-DRB-16 is protective for the development of PsA with coefficients estimated as -0.9284 for the ALASSO and -0.9317 for the SCAD penalty functions. When the 5-fold CV statistic is used to select the tuning parameter, we find HLA-DRB1-10 and HLA-DRB-16 are both identified using ALASSO (coefficient estimates of -0.7144 and -0.9749 respectively) and SCAD (-0.7160 and -0.9771 respectively). See Figure 2.5 for plots of the cross-validation statistics and traces of the parameter estimates.

2.6.4 Discussion

In this section we develop methods for variable selection based on truncated and interval-censored data. Increasingly often scientists have two or more large datasets available with each providing information on chronic disease processes. Such datasets typically have different criteria for the inclusion of patients which routinely leads to truncated data. A natural question arises regarding the value of left and right-truncated data in identifying genetic risks for disease. For right-truncated data all event times may be known with some uncertainty since they may be retrospectively recorded with error. For left-truncated data the key factor is the duration of follow-up and the number of individuals going on to experience the event of interest. A third type of data may be available to enhance efficiency in estimation and identification of key genetic risk factors. Cross-sectional examination of individuals with diabetes with known onset times can be useful if their retinopathy status can be determined along with genetic samples.

Appendix 2.A Supplementary Simulation Studies

Here we conduct simulation studies with a relatively small number of covariates ($p = 10$). The generating procedure for the normally distributed covariates is the same as described in Section 2.3 with $p = 100$ multivariate normal covariates. We set $\beta_j = 0.5$, $j = 1, 2, 9, 10$ and $\beta_j = 0$, $j = 3, \ldots, 6$. In Table 2.A.1, we report the results of applying the penalized expectation-maximization algorithm (P-EM) to proportional hazards models with exponential (EXP) and piecewise constant baseline hazards with four pieces (PWC-4). We also report corresponding results following mid-point imputation when the resulting data are treated as right-censored (MID). Traditional methods of variable selection based on forward selection and backward elimination are also considered based on the correct parametric Weibull regression model.

For the case of correlated binary covariates, the data are generated using a series of conditional binary probability mass functions as described by Preisser et al. (2002). We set the marginal probabilities such that $E(X_{ij}) = 0.05$, $j = 1, \ldots, 5$ and $E(X_{ij}) = 0.20$, $j = 6, \ldots, 10$, using a 10×10 correlation matrix with entry $\text{corr}(X_{ij}, X_{ik}) = \rho^{|j-k|}$, where $\rho = 0.3$ or 0.6. The coefficients in the proportional hazards model are set to $\beta_j = 1$ for $j = 1, 2, 9, 10$ and $\beta_j = 0$, $j = 3, \ldots, 6$. The analyses are the same as those used for the multivariate normal covariates; and Table 2.A.2 shows the results which is analogous to Table 2.A.1.

When comparing the results between the midpoint imputed and interval-censored datasets with the PWC-4 model in Table 2.A.1 and Table 2.A.2, there is generally a comparable ability to detecting important covariates (TP) and number of false positive (FP) selections, but the proposed P-EM algorithm leads to lower MSE. The results from traditional variable selection methods also feature high mean squared errors and slightly higher FP

values.

Figure 2.A.1 contains box plots of the empirical estimation errors $(\widehat{\beta}_j - \beta_j)$ for four of the ten coefficients (β_1 and β_2 (both equal to 1) and β_3 and β_5 (both equal to zero)) when data are simulated with $\kappa = 1.25$, $\mu = 10$ and $\rho = 0.3$. We report on results for an exponential and piecewise constant baseline hazard, for datasets featuring by mid-point imputation (MID), interval-censoring (P-EM), and for the case where the actual event time is used, subject only to right-censoring (RC). The performance of the piecewise constant model is generally better than the exponential model since $\kappa \neq 1$, and for this hazard function, the P-EM algorithm leads to performance which is more like the analysis using the right-censored (RC) failure time; the latter analysis is only possible in a simulation study such as this where the interval-censored time is actually known.

			$\rho = 0.3$						$\rho = 0.6$					
			$\mu = 10$			$\mu = 20$			$\mu = 10$			$\mu = 20$		
Model	Penalty	Method	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)
								Shape parameter: $\kappa = 1$						
EXP	LASSO	P-EM	4.00	3.08	0.021 (0.012)	4.00	2.51	0.019 (0.013)	4.00	2.15	0.020 (0.014)	4.00	2.48	0.020 (0.012)
		MID	4.00	2.73	0.077 (0.029)	4.00	2.38	0.033 (0.018)	4.00	2.09	0.117 (0.037)	4.00	2.59	0.044 (0.029)
	ALASSO	P-EM	4.00	0.52	0.012 (0.013)	4.00	0.30	0.012 (0.011)	4.00	0.45	0.014 (0.012)	4.00	0.68	0.012 (0.013)
		MID	4.00	0.45	0.048 (0.023)	4.00	0.31	0.018 (0.013)	4.00	0.53	0.084 (0.029)	4.00	0.66	0.029 (0.018)
	SCAD	P-EM	4.00	0.51	0.010 (0.013)	4.00	0.40	0.012 (0.012)	4.00	0.33	0.013 (0.011)	4.00	0.61	0.011 (0.012)
		MID	4.00	0.38	0.048 (0.023)	4.00	0.35	0.018 (0.012)	4.00	0.44	0.082 (0.029)	4.00	0.55	0.028 (0.019)
PWC-4	LASSO	P-EM	4.00	3.05	0.026 (0.017)	4.00	2.54	0.020 (0.015)	4.00	2.10	0.027 (0.018)	4.00	2.38	0.022 (0.015)
		MID	4.00	2.84	0.057 (0.030)	4.00	2.50	0.029 (0.019)	4.00	2.19	0.085 (0.038)	4.00	2.55	0.037 (0.024)
	ALASSO	P-EM	4.00	0.45	0.012 (0.012)	4.00	0.21	0.013 (0.014)	4.00	0.45	0.015 (0.013)	4.00	0.59	0.013 (0.014)
		MID	4.00	0.29	0.029 (0.021)	4.00	0.34	0.015 (0.013)	4.00	0.38	0.052 (0.029)	4.00	0.56	0.019 (0.017)
	SCAD	P-EM	4.00	0.45	0.012 (0.013)	4.00	0.31	0.014 (0.014)	4.00	0.34	0.015 (0.013)	4.00	0.56	0.012 (0.013)
		MID	4.00	0.34	0.029 (0.021)	4.00	0.46	0.015 (0.012)	4.00	0.38	0.052 (0.029)	4.00	0.59	0.020 (0.017)
FORWARD			4.00	0.57	0.014 (0.012)	4.00	0.46	0.017 (0.011)	4.00	0.45	0.017 (0.012)	4.00	0.47	0.014 (0.011)
BACKWARD			4.00	0.65	0.014 (0.012)	4.00	0.48	0.017 (0.011)	4.00	0.60	0.018 (0.012)	4.00	0.65	0.016 (0.011)
								Shape parameter: $\kappa = 1.25$						
EXP	LASSO	P-EM	4.00	2.80	0.057 (0.022)	4.00	2.53	0.057 (0.023)	4.00	2.26	0.063 (0.026)	4.00	2.50	0.064 (0.022)
		MID	4.00	2.68	0.113 (0.034)	4.00	2.47	0.076 (0.025)	4.00	2.10	0.157 (0.039)	4.00	2.39	0.094 (0.029)
	ALASSO	P-EM	4.00	0.47	0.030 (0.017)	4.00	0.33	0.032 (0.017)	4.00	0.48	0.038 (0.021)	4.00	0.72	0.041 (0.018)
		MID	4.00	0.38	0.082 (0.028)	4.00	0.35	0.047 (0.020)	4.00	0.57	0.123 (0.034)	4.00	0.55	0.068 (0.023)
	SCAD	P-EM	4.00	0.59	0.030 (0.018)	4.00	0.32	0.032 (0.017)	4.00	0.35	0.039 (0.021)	4.00	0.52	0.040 (0.018)
		MID	4.00	0.60	0.082 (0.028)	4.00	0.42	0.049 (0.020)	4.00	0.47	0.123 (0.033)	4.00	0.53	0.067 (0.023)
PWC-4	LASSO	P-EM	4.00	2.94	0.025 (0.015)	4.00	2.52	0.021 (0.016)	4.00	2.26	0.023 (0.017)	4.00	2.45	0.022 (0.014)
		MID	4.00	3.04	0.043 (0.027)	4.00	2.78	0.028 (0.017)	4.00	2.27	0.066 (0.033)	4.00	2.54	0.031 (0.022)
	ALASSO	P-EM	4.00	0.42	0.010 (0.012)	4.00	0.27	0.012 (0.012)	4.00	0.44	0.015 (0.014)	4.00	0.58	0.011 (0.013)
		MID	4.00	0.46	0.022 (0.020)	4.00	0.30	0.014 (0.011)	4.00	0.29	0.038 (0.023)	4.00	0.53	0.017 (0.014)
	SCAD	P-EM	4.00	0.48	0.010 (0.012)	4.00	0.28	0.013 (0.012)	4.00	0.41	0.016 (0.013)	4.00	0.55	0.011 (0.013)
		MID	4.00	0.41	0.022 (0.020)	4.00	0.41	0.015 (0.011)	4.00	0.32	0.038 (0.023)	4.00	0.52	0.017 (0.015)
FORWARD			4.00	0.53	0.060 (0.022)	4.00	0.51	0.060 (0.022)	4.00	0.47	0.076 (0.028)	4.00	0.61	0.073 (0.024)
BACKWARD			4.00	0.69	0.061 (0.023)	4.00	0.51	0.060 (0.022)	4.00	0.62	0.078 (0.028)	4.00	0.72	0.072 (0.024)

Table 2.A.1: Empirical results for interval-censored data with normally distributed covariates ($p = 10$, $E(X_{ij}) = 0$, $Var(X_{ij}) = 1$ and $\text{corr}(X_{ij}, X_{ik}) = \rho^{|j-k|}$) summarizing the number of correctly (TP) and incorrectly (FP) selected variables along with the median and the standard deviation (SD) of the mean squared error (MSE); P-EM denotes the analyses based on the proposed penalized EM method and MID denotes an analysis based on a pseudo-data set obtained by mid-point imputation; the tuning parameter is selected by five-fold cross validation.

			$\rho = 0.3$						$\rho = 0.6$					
			$\mu = 10$			$\mu = 20$			$\mu = 10$			$\mu = 20$		
Model	Penalty	Method	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)
			Shape parameter: $\kappa = 1$											
EXP	LASSO	P-EM	4.00	2.48	0.293 (0.230)	3.99	2.77	0.263 (0.225)	3.99	2.43	0.292 (0.227)	4.00	2.12	0.256 (0.176)
		MID	3.99	2.53	0.826 (0.353)	3.99	2.56	0.451 (0.217)	3.96	2.35	1.102 (0.481)	4.00	2.02	0.506 (0.276)
	ALASSO	P-EM	3.82	0.77	0.250 (0.486)	3.93	0.90	0.208 (0.313)	3.84	1.44	0.287 (0.300)	3.93	0.79	0.162 (0.409)
		MID	3.79	0.73	0.611 (0.451)	3.92	0.88	0.263 (0.300)	3.64	0.83	0.968 (0.573)	3.88	0.63	0.317 (0.276)
	SCAD	P-EM	3.84	0.62	0.200 (0.465)	3.91	0.79	0.217 (0.402)	3.71	1.16	0.337 (0.339)	3.89	0.58	0.217 (0.300)
		MID	3.78	0.45	0.573 (0.459)	3.94	0.77	0.261 (0.280)	3.61	0.79	0.961 (0.578)	3.84	0.58	0.331 (0.283)
PWC-4	LASSO	P-EM	4.00	2.34	0.314 (0.239)	4.00	2.47	0.262 (0.217)	3.99	2.21	0.305 (0.238)	4.00	1.91	0.265 (0.193)
		MID	4.00	2.43	0.602 (0.320)	3.99	2.48	0.363 (0.208)	3.97	2.35	0.844 (0.440)	4.00	1.97	0.381 (0.254)
	ALASSO	P-EM	3.79	0.65	0.276 (0.567)	3.92	0.82	0.223 (0.328)	3.80	1.13	0.309 (0.428)	3.88	0.73	0.215 (0.426)
		MID	3.72	0.48	0.343 (0.635)	3.92	0.61	0.218 (0.297)	3.60	0.84	0.729 (0.735)	3.88	0.49	0.220 (0.397)
	SCAD	P-EM	3.85	0.69	0.210 (0.481)	3.95	0.89	0.213 (0.301)	3.75	1.18	0.314 (0.444)	3.89	0.69	0.236 (0.310)
		MID	3.74	0.52	0.351 (0.599)	3.92	0.73	0.229 (0.295)	3.63	0.87	0.720 (0.638)	3.88	0.43	0.221 (0.275)
FORWARD			3.99	0.63	0.216 (0.318)	3.97	0.67	0.227 (0.272)	3.79	0.59	0.269 (0.335)	3.91	0.56	0.202 (0.339)
BACKWARD			3.99	0.64	0.216 (0.315)	3.97	0.69	0.227 (0.275)	3.79	0.78	0.275 (0.332)	3.91	0.81	0.223 (0.334)
			Shape parameter: $\kappa = 1.25$											
EXP	LASSO	P-EM	4.00	2.36	0.544 (0.254)	4.00	2.70	0.443 (0.222)	3.98	2.34	0.536 (0.275)	3.99	2.04	0.508 (0.217)
		MID	4.00	2.32	0.994 (0.303)	4.00	2.56	0.613 (0.221)	3.96	2.17	1.303 (0.412)	3.99	2.19	0.771 (0.244)
	ALASSO	P-EM	3.89	0.77	0.296 (0.468)	3.93	0.77	0.252 (0.275)	3.82	1.19	0.383 (0.281)	3.90	0.74	0.312 (0.219)
		MID	3.91	0.73	0.721 (0.343)	3.97	0.93	0.423 (0.238)	3.76	0.82	1.036 (0.443)	3.92	0.72	0.553 (0.238)
	SCAD	P-EM	3.88	0.54	0.270 (0.426)	3.94	0.64	0.247 (0.265)	3.75	0.97	0.428 (0.299)	3.88	0.70	0.314 (0.244)
		MID	3.87	0.47	0.718 (0.317)	3.96	0.75	0.425 (0.243)	3.68	0.68	1.049 (0.385)	3.83	0.39	0.550 (0.291)
PWC-4	LASSO	P-EM	3.99	2.15	0.284 (0.240)	3.99	2.32	0.249 (0.197)	3.98	2.18	0.308 (0.222)	3.99	1.86	0.245 (0.184)
		MID	4.00	2.51	0.489 (0.251)	4.00	2.70	0.304 (0.181)	3.95	2.28	0.617 (0.383)	3.99	2.00	0.332 (0.217)
	ALASSO	P-EM	3.83	0.56	0.173 (0.561)	3.94	0.62	0.153 (0.307)	3.83	1.17	0.271 (0.293)	3.91	0.76	0.182 (0.267)
		MID	3.88	0.57	0.279 (0.317)	3.95	0.64	0.159 (0.267)	3.69	0.70	0.480 (0.568)	3.88	0.72	0.210 (0.247)
	SCAD	P-EM	3.84	0.55	0.165 (0.514)	3.95	0.74	0.156 (0.299)	3.81	1.19	0.282 (0.292)	3.88	0.69	0.148 (0.280)
		MID	3.85	0.62	0.288 (0.336)	3.94	0.57	0.159 (0.271)	3.70	0.66	0.480 (0.437)	3.88	0.63	0.189 (0.255)
FORWARD			3.96	0.61	0.326 (0.234)	3.98	0.70	0.303 (0.197)	3.80	0.61	0.372 (0.297)	3.92	0.59	0.345 (0.213)
BACKWARD			3.96	0.65	0.326 (0.233)	3.98	0.73	0.303 (0.198)	3.80	0.76	0.395 (0.289)	3.91	0.82	0.383 (0.233)

Table 2.A.2: Empirical results for interval-censored data with correlated binary covariates ($p = 10$, $E(X_{ij}) = 0.2$ and $\text{corr}(X_{ij}, X_{ik}) = \rho^{|j-k|}$) summarizing the number of correctly (TP) and incorrectly (FP) selected variables along with the median and the standard deviation (SD) of the mean squared error (MSE); P-EM denotes the analyses based on the proposed penalized EM method and MID denotes an analysis based on a pseudo-data set obtained by mid-point imputation; the tuning parameter is selected by five-fold cross validation.

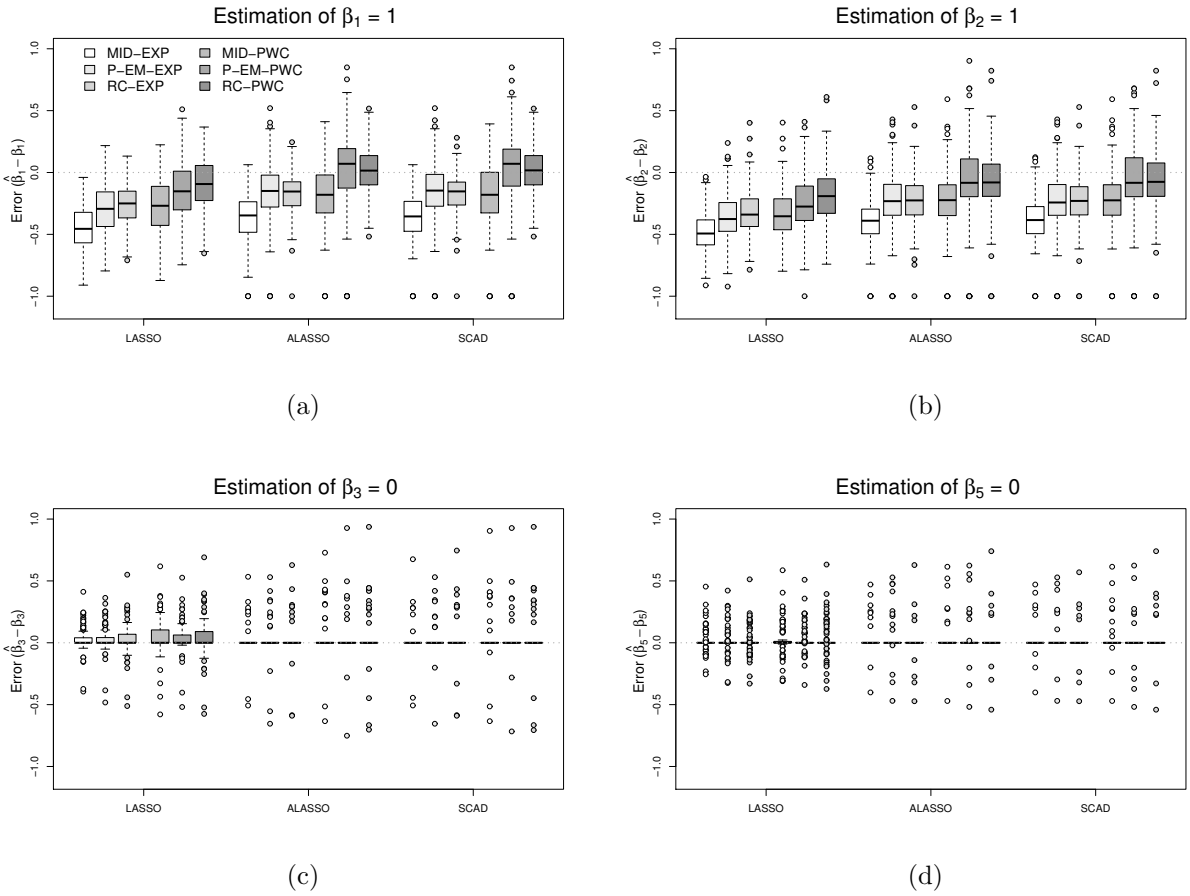


Figure 2.A.1: Box plots of the error for the estimated regression coefficients $\widehat{\beta}_k - \beta_k$, $k = 1, 2, 3, 5$, for each penalty function for datasets with correlated binary covariates ($p = 10$) with $\kappa = 1.25$, $\mu = 10$, $\rho = 0.3$.

Appendix 2.B Comparison of Methods for Choosing the Optimal Tuning Parameter

The selection of the tuning parameter λ is an important step in analyses based on penalized likelihood; when $\lambda = \infty$, none of the variables will be selected and when $\lambda = 0$, all of the variables will be selected in the usual fashion. Classical model selection methods are often based on the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) and more recent strategies have been based on cross-validation (CV) and generalized cross-validation (GCV). The traditional G-fold CV statistic is defined as

$$\widehat{CV}(\lambda) = \sum_{g=1}^{G} \left[\log L(\widehat{\boldsymbol{\theta}}_{-g}(\lambda)) - \log L_{-g}(\widehat{\boldsymbol{\theta}}_{-g}(\lambda)) \right]$$

where L_{-g} is the likelihood for the gth training dataset and $\widehat{\boldsymbol{\theta}}_{-g}(\lambda)$ is the estimate for the gth training data, obtained through the EM algorithm; the optimal λ maximizes $\widehat{CV}(\lambda)$.

Bradic et al. (2011) mentioned that the measure of information contained in the full Cox partial likelihood is biased with respect to the number of nonzero elements and proper normalization is required. They proposed a sparse approximation to the generalized cross-validation statistic (SGCV) as

$$\widehat{SGCV}(\lambda) = \sum_{g=1}^{G} \left[\frac{\log L(\widehat{\boldsymbol{\theta}}_{-g}(\lambda))}{m\left(1 - \widehat{s}_{-g}(\lambda)/m\right)^2} - \frac{\log L_{-g}(\widehat{\boldsymbol{\theta}}_{-g}(\lambda))}{m_{-g}\left(1 - \widehat{s}_{-g}(\lambda)/m_{-g}\right)^2} \right]$$

where m_{-g} is the sample size of the gth training dataset and $\widehat{s}_{-g}(\lambda)$ is the number of non-zero coefficients. The optimal λ minimizes $\widehat{SGCV}(\lambda)$.

Here we compare three methods of selecting tuning parameters: cross-validation (CV), Bayesian information criterion (BIC) and sparse generalized cross-validation (SGCV). Table 2.B.1 shows the results of comparisons of three methods for proportional hazards models

with a piecewise constant baseline hazards with four pieces (PWC-4) for datasets with correlated binary covariates of dimension $p = 10$; Table 2.B.2 shows the corresponding results for datasets with multivariate normal covariates of dimension $p = 100$.

From these two tables, we see that for the LASSO penalty, SGCV shows some improvements in terms of a smaller number of incorrectly selected variables (FP); however, it also results in a smaller number of correctly selected variables (TP) and a larger mean squared error (MSE).

Compared with SGCV, both BIC and CV show good performance in terms of selecting tuning parameters for the ALASSO and SCAD penalties; BIC shows a smaller number of incorrectly selected variables (FP) than CV for LASSO penalty. Since, the R package `glmnet` uses cross-validation, we report the corresponding implementation of our algorithm using cross-validation to select tuning parameter.

		$\rho = 0.3$						$\rho = 0.6$					
		$\mu = 10$			$\mu = 20$			$\mu = 10$			$\mu = 20$		
Penalty	Method	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)
							Shape parameter: $\kappa = 1$						
LASSO	CV	4.00	2.34	0.314 (0.239)	4.00	2.47	0.262 (0.217)	3.99	2.21	0.305 (0.238)	4.00	1.91	0.265 (0.193)
	BIC	3.99	0.86	0.447 (0.305)	3.98	0.99	0.329 (0.251)	3.94	0.96	0.406 (0.315)	3.99	0.92	0.287 (0.231)
	SGCV	2.90	0.49	3.025 (1.411)	3.09	0.93	1.125 (1.463)	1.23	0.29	6.543 (2.290)	2.81	0.56	3.709 (1.982)
ALASSO	CV	3.79	0.65	0.276 (0.567)	3.92	0.82	0.223 (0.328)	3.80	1.13	0.309 (0.428)	3.88	0.73	0.215 (0.426)
	BIC	3.75	0.08	0.188 (0.502)	3.85	0.08	0.150 (0.373)	3.53	0.07	0.576 (0.355)	3.70	0.07	0.207 (0.377)
	SGCV	1.74	0.03	2.659 (1.564)	1.39	0.06	3.511 (1.591)	1.86	0.06	3.272 (2.079)	1.96	0.10	3.223 (2.006)
SCAD	CV	3.85	0.69	0.210 (0.481)	3.95	0.89	0.213 (0.301)	3.75	1.18	0.314 (0.444)	3.89	0.69	0.236 (0.310)
	BIC	3.73	0.06	0.176 (0.505)	3.85	0.07	0.148 (0.373)	3.48	0.04	0.670 (0.358)	3.65	0.05	0.237 (0.366)
	SGCV	1.50	0.00	3.516 (1.121)	1.60	0.02	3.511 (1.252)	1.44	0.02	3.914 (1.610)	1.38	0.03	3.908 (1.773)
							Shape parameter: $\kappa = 1.25$						
LASSO	CV	3.99	2.15	0.284 (0.240)	3.99	2.32	0.249 (0.197)	3.98	2.18	0.308 (0.222)	3.99	1.86	0.245 (0.184)
	BIC	3.99	0.82	0.364 (0.320)	3.98	0.91	0.303 (0.257)	3.94	0.68	0.371 (0.308)	3.98	0.83	0.269 (0.243)
	SGCV	2.95	0.30	2.235 (1.409)	3.09	0.85	1.157 (1.475)	2.97	0.75	2.111 (1.905)	2.97	0.71	1.756 (1.952)
ALASSO	CV	3.83	0.56	0.173 (0.561)	3.94	0.62	0.153 (0.307)	3.83	1.17	0.271 (0.293)	3.91	0.76	0.182 (0.267)
	BIC	3.84	0.05	0.127 (0.361)	3.87	0.14	0.141 (0.341)	3.49	0.03	0.653 (0.349)	3.67	0.02	0.166 (0.306)
	SGCV	1.83	0.02	2.624 (1.565)	1.44	0.02	3.511 (1.466)	1.33	0.08	3.846 (2.126)	2.29	0.07	3.203 (2.001)
SCAD	CV	3.84	0.55	0.165 (0.514)	3.95	0.74	0.156 (0.299)	3.81	1.19	0.282 (0.292)	3.88	0.69	0.148 (0.280)
	BIC	3.81	0.06	0.130 (0.380)	3.86	0.14	0.141 (0.356)	3.48	0.02	0.656 (0.402)	3.66	0.01	0.182 (0.308)
	SGCV	1.53	0.02	3.518 (1.314)	1.58	0.02	3.511 (1.326)	1.41	0.00	3.914 (1.564)	1.42	0.02	3.903 (1.557)

Table 2.B.1: Comparison of three methods of choosing tuning parameter: cross-validation (CV), Bayesian information criterion (BIC) and sparse generalized cross-validation (SGCV). Analyses were based on interval-censored responses with correlated binary covariates ($p = 10$) by using proportional hazards models with a piecewise constant baseline hazards with four pieces (PWC-4) and results are summarized in terms of the number of correctly (TP) and incorrectly (FP) selected variables and the median and standard deviation of the mean squared error (MSE).

		$\mu = 10$			$\mu = 20$		
Penalty	Method	TP (10)	FP (90)	MSE (SD)	TP (10)	FP (90)	MSE (SD)
				Shape parameter: $\kappa = 1$			
LASSO	CV	10.00	14.80	0.312 (0.126)	10.00	14.83	0.261 (0.105)
	BIC	10.00	3.34	0.624 (0.199)	10.00	3.94	0.512 (0.184)
	SGCV	9.72	5.36	1.405 (0.823)	9.79	5.12	1.246 (0.692)
ALASSO	CV	10.00	0.12	0.057 (0.047)	10.00	0.07	0.047 (0.040)
	BIC	10.00	0.72	0.084 (0.072)	10.00	0.84	0.076 (0.057)
	SGCV	8.25	43.21	1.178 (1.329)	8.55	46.99	0.992 (1.011)
SCAD	CV	9.98	0.36	0.059 (0.073)	9.99	0.24	0.050 (0.048)
	BIC	10.00	0.84	0.082 (0.081)	10.00	0.79	0.068 (0.064)
	SGCV	9.55	58.93	1.275 (0.690)	9.51	53.23	0.940 (0.784)
				Shape parameter: $\kappa = 1.25$			
LASSO	CV	10.00	14.88	0.291 (0.118)	10.00	14.13	0.245 (0.109)
	BIC	10.00	3.37	0.604 (0.184)	10.00	3.78	0.501 (0.164)
	SGCV	9.81	0.96	1.277 (0.707)	9.64	2.70	1.227 (0.877)
ALASSO	CV	9.99	0.23	0.055 (0.050)	10.00	0.08	0.045 (0.031)
	BIC	10.00	0.59	0.068 (0.075)	10.00	0.90	0.071 (0.047)
	SGCV	9.54	62.70	1.024 (0.766)	7.14	29.37	0.983 (1.501)
SCAD	CV	9.98	0.29	0.055 (0.052)	9.99	0.13	0.044 (0.036)
	BIC	10.00	0.62	0.070 (0.085)	10.00	0.90	0.069 (0.058)
	SGCV	7.45	49.44	1.207 (1.987)	8.93	35.06	0.716 (0.735)

Table 2.B.2: Comparison of three methods of choosing tuning parameter: cross-validation (CV), Bayesian information criterion (BIC) and sparse generalized cross-validation (SGCV). Analyses were based on interval-censored responses with multivariate normal covariates ($p = 100$) by using proportional hazards models with a piecewise constant baseline hazards with four pieces (PWC-4) and results are summarized in terms of the number of correctly (TP) and incorrectly (FP) selected variables and the median and standard deviation of the mean squared error (MSE).

Appendix 2.C Variance Estimation

It is difficult to obtain an accurate estimate of the standard errors of the penalized estimator since the estimate is a non-linear and non-differentiable function of the responses, even for a fixed tuning parameter. Moreover, the sampling distribution would have a point mass at zero and it is even unclear how one could use a standard error for inference. Some authors, however, estimate the variance by using approximations or the bootstrap and we consider this here. This work was motivated by a referee comment in a manuscript (Wu and Cook, 2015).

For the LASSO penalty, Tibshirani (1996, 1997) suggested estimating standard errors using either the bootstrap with either a fixed or an unfixed tuning parameter, or using an approximate form derived from ridge regression. For the SCAD penalty, Fan and Li (2001) suggested that for moderate sample sizes, a sandwich-type variance formula derived from a local quadratic approximation (LQA) could be used for the covariance matrix, with modifications for large sample sizes. For the adaptive LASSO penalty, Zou (2006) also used a LQA sandwich formula to approximate the variance of the estimators from penalized likelihood.

In the main paper, we propose an approach to variable selection for interval-censored failure times via a piecewise exponential model; it is not easy to derive an approximate approach to estimate standard errors. Therefore, we have employed a bootstrap approach to calculate standard errors of the penalized estimators. We draw a random sample $\boldsymbol{D}^*$ of size $m = 500$ with replacement from the original dataset $\boldsymbol{D}$ and we can obtain the penalized estimates $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)$ from $\boldsymbol{D}^*$ by using the proposed method with tuning parameter fixed at the optimal value that was determined from the original dataset $\boldsymbol{D}$. We repeat this process 500 times and get 500 bootstrap penalized estimates

$\boldsymbol{\beta}^{*(1)}, \ldots, \boldsymbol{\beta}^{*(500)}$, so the bootstrap standard errors of the penalized estimators will be given by $\text{SE}(\beta_1^{*(1)}, \ldots, \beta_1^{*(500)}), \ldots, \text{SE}(\beta_p^{*(1)}, \ldots, \beta_p^{*(500)})$. Table 2.C.1 shows the empirical biases, the average of the bootstrap standard errors, the empirical standard errors for the simulated datasets with $p = 10$, $\kappa = 1.25$, $\mu = 10$, $\rho = 0.3$ for both multivariate normal covariates and multivariate binary covariates. We can see that for the non-zero coefficients $(\beta_1, \beta_2, \beta_9, \beta_{10})$, the ASE and ESE agree well; for the zero coefficients, the ASE tends to be bigger than the ESE. We note that although we can calculate the standard errors based on the bootstrap or approximate approaches, it remains challenging to conceive how one would construct a confidence interval or compute a $p-$value based on a standard Wald-based pivotal or test statistic.

Penalty		β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
					Multivariate Normal Covariate						
LASSO	EBIAS	-0.042	-0.043	0.008	-0.004	-0.002	0.000	0.005	-0.004	-0.052	-0.048
	ASE	0.057	0.057	0.035	0.036	0.036	0.035	0.036	0.036	0.058	0.057
	ESE	0.058	0.050	0.028	0.031	0.035	0.029	0.037	0.034	0.053	0.048
ALASSO	EBIAS	0.004	0.003	0.003	-0.001	0.001	0.000	0.002	-0.003	-0.006	-0.003
	ASE	0.061	0.062	0.043	0.045	0.046	0.044	0.045	0.045	0.062	0.060
	ESE	0.059	0.052	0.018	0.023	0.027	0.024	0.034	0.028	0.052	0.047
SCAD	EBIAS	0.004	0.002	0.005	-0.001	0.001	0.001	0.001	-0.003	-0.006	-0.003
	ASE	0.061	0.062	0.047	0.047	0.048	0.047	0.047	0.048	0.063	0.061
	ESE	0.059	0.052	0.020	0.021	0.029	0.020	0.036	0.030	0.052	0.048
					Multivariate Binary Covariate						
LASSO	EBIAS	-0.155	-0.231	0.033	0.032	0.020	0.008	0.003	0.005	-0.097	-0.100
	ASE	0.240	0.249	0.142	0.135	0.139	0.069	0.073	0.075	0.132	0.133
	ESE	0.258	0.258	0.122	0.118	0.100	0.059	0.057	0.075	0.132	0.139
ALASSO	EBIAS	-0.011	-0.071	0.018	0.036	0.016	0.005	0.001	-0.001	0.012	0.010
	ASE	0.259	0.273	0.182	0.170	0.175	0.094	0.096	0.095	0.143	0.141
	ESE	0.392	0.372	0.148	0.146	0.127	0.050	0.061	0.084	0.140	0.141
SCAD	EBIAS	-0.002	-0.071	0.018	0.031	0.007	0.005	-0.002	-0.003	0.013	0.009
	ASE	0.259	0.272	0.182	0.179	0.181	0.094	0.099	0.099	0.142	0.141
	ESE	0.381	0.366	0.145	0.139	0.119	0.042	0.053	0.083	0.142	0.141

Table 2.C.1: Variance estimation by bootstrap for the simulated dataset with multivariate normal covariates and multivariate binary covariates for $\kappa = 1.25$, $\mu = 10$, $\rho = 0.3$.

Chapter 3

Assessing the Accuracy of Predictive Models with Interval-Censored Data

3.1 Introduction

3.1.1 Overview

In the context of time to event data, we obtain flexible prediction models and often evaluate their predictive value on the same set of data, or a validation data. The purpose of assessing the predictive accuracy of a regression model is to establish whether a prognostic model can be used to reliably predict patients' event status and provide a basis for clinical decision making. Predictive accuracy can also be used as a strategy for model selection.

There has been a lot of research conducted focusing on prediction with time to event data where the event times are subject to right-censoring. Various ways and aspects to assess the predictive performance of a statistical model have been studied by many authors;

two typical outcomes of interest are the event time and patients' status at a particular time. In health-related research, scientists are often interested in predicting patients' future status based on covariates; for example, the presence of progression in 2 years after the onset of disease for a patient with age of 60 at clinical entry.

Common approaches to quantify the overall performance of the prediction model are using the measures such as the explained variation, the Brier score and the loss functions. In order to determine whether a predictor $\widehat{Y}$ predicts well of Y, we consider a loss function, which measures the distance between the predicted and true values. The loss function, when averaged over all possible values of the data yields a measure of the prediction error. The absolute loss function is $L(Y, \widehat{Y}) = |Y - \widehat{Y}|$ and the squared loss function is $L(Y, \widehat{Y}) = (Y - \widehat{Y})^2$. The difficulty in assessing predictive accuracy due to censoring has been studied by several authors. Korn and Simon (1990) proposed a bounded loss function to be used for predicting survival time. Inverse probability weighting (IPW) approach used by Graf et al. (1999), Hothorn et al. (2006), Gerds and Schumacher (2006), Lawless and Yuan (2010) to deal with censored outcomes.

When the response is a binary indicator of the survival status at a specific time t_0, a key component of assessing the predictive performance is the ability of correctly classifying individuals with respect to their status at time t_0. The features of interest include the sensitivity $P(\widehat{Y} = 1|Y = 1)$ and specificity $P(\widehat{Y} = 0|Y = 0)$. The discriminative ability can then be quantified through construction of a receiver operating characteristic (ROC) curve, which is obtained by plotting the true positive rate (sensitivity) against the false positive rate (1 - specificity) for different thresholds. Akritas (1994) proposed an estimator based on a nearest neighbor algorithm for the bivariate distribution function $P(\widehat{Y}, Y)$, which can guarantee the monotonicity in terms of sensitivity and specificity; an alternative simple estimator which does not guarantee monotonicity (Heagerty et al., 2000) based on

the sensitivity and specificity involves the Kaplan-Meier estimate. Yuan (2008) discussed an estimator for right-censored data which does guarantee monotonicity based on inverse probability of censoring weights.

3.1.2 Estimating Prediction Error

Let T_i and C_i denote the event time of interest and the right censoring time for individual i respectively; we observe $\min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. If X_i denotes a $p \times 1$ covariate vector for individual i, then the data from a training sample of n independent individuals are denoted by $D = \{(\min(T_i, C_i), \delta_i, X_i), i = 1, \ldots, n\}$. A corresponding validation dataset of $m - n$ independent individuals is denoted by $\bar{D} = \{(\min(T_j, C_j), \delta_j, X_j), j = n+1, \ldots, m\}$.

If t_0 denotes a specific landmark time, one can define a binary status indicator $Y = I(T > t_0)$, which indicates that an individual is event-free at t_0; we let $\widehat{Y}(X; \theta)$ denote a prediction for Y which is based on a model for $Y|X$ indexed by θ. To examine the predictive accuracy of such models with a binary response, traditional methods often involve a summary statistic reflecting overall predictive performance such as the mean squared error (Efron, 1983) or a concordance statistic for discriminative ability such as the area under a receiver operating characteristic curve (Hanley and McNeil, 1982).

Overall Predictive Performance

The prediction error based on a squared error loss function is defined as

$$\text{PE} = E[\{Y - \widehat{Y}(X; \widehat{\theta})\}^2] \,, \tag{3.1}$$

where $\widehat{\theta} = \widehat{\theta}(D)$ is the estimated parameter of the prediction model. The expectation in (3.1) is taken with respect to (i) the response variable Y, (ii) the covariate vector X, and

(iii) the training data D. The distribution of the covariates X is typically unknown and it is undesirable for the prediction error to be dependent on specification of a possibly high dimensional covariate distribution, so the empirical distribution of X is usually used for the expectation with respect to X to estimate the prediction error.

The optimal predictor is the one that minimizes the predictor error. For (3.1), the optimal predictor is the conditional probability of survival to t_0 given the covariate which is $\widehat{Y}(X;\widehat{\theta}) = P(T > t_0|X;\widehat{\theta})$. If we focus on the predictor with the same support as Y then we use

$$\widehat{Y}(X;\widehat{\theta}) = I(P(T > t_0|X;\widehat{\theta}) > c) \,, \tag{3.2}$$

which is the optimal binary predictor when the threshold $c = 0.5$ is used; we focus on the binary predictor from here on.

With a validation dataset, the prediction error can be estimated empirically by

$$\widehat{\mathrm{PE}}(t_0) = \frac{1}{m-n} \sum_{j=n+1}^{m} \left\{ Y_j - \widehat{Y}_j(X_j;\widehat{\theta}) \right\} \,. \tag{3.3}$$

When there is no validation dataset, there are three broad approaches for estimating the prediction error: an apparent loss-based estimator, an estimator based on cross-validation, and a model-based estimator. We define these three approaches in what follows.

The apparent loss error is defined as

$$\widehat{\mathrm{PE}}(t_0) = \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - \widehat{Y}_i(X_i;\widehat{\theta}) \right\}^2 \,. \tag{3.4}$$

When a validation dataset is not available, a naive estimate of the apparent loss based on the training data tends to underestimate the true loss because Y_i and $\widehat{Y}_i(X_i;\widehat{\theta})$ are positively correlated (Efron, 1986). That is, the apparent loss approach compares predictions based on the same observations as are used for the development of the prediction model;

for flexible prediction rules, this negative bias could be a serious issue (Gerds and Schumacher, 2007). Many approaches have been developed to estimate this bias and obtain bias-corrected measures of the prediction error in different regression settings and with different loss functions (Mallows, 1973; Efron, 1983, 1986, 2004).

Cross-validation is a widely used technique to estimate prediction error in the absence of a validation sample. In this setting the data D is split into G subsamples $\mathcal{S}_1, \ldots, \mathcal{S}_G$; we refer to $\mathcal{S}_g$ and $\mathcal{S} - \mathcal{S}_g$ as the gth test and training sets, $g = 1, \ldots, G$. The G-fold cross-validation estimate for the prediction error is defined as

$$\widehat{\mathrm{PE}}(t_0) = \frac{1}{n} \sum_{g=1}^{G} \sum_{i \in \mathcal{S}_g} \left\{ Y_i - \widehat{Y}_i(X_i; \widehat{\theta}_{-g}) \right\}^2 , \tag{3.5}$$

where $\widehat{\theta}_{-g}$ is the estimate for the gth training data. A variant of this approach is to use *bootstrap cross-validation*, wherein instead of splitting the training data into distinct subsamples, B bootstrap samples $D_1^*, \ldots, D_B^*$ of size n are drawn *with replacement* from the original data D. The estimate of the prediction error is then

$$\widehat{\mathrm{PE}}(t_0) = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{n_b} \sum_{i: i \notin D_b^*} \left\{ Y_i - \widehat{Y}_i(X_i; \widehat{\theta}_b^*) \right\}^2 , \tag{3.6}$$

where $\widehat{\theta}_b^*$ is the estimate for the bth bootstrap sample and n_b denotes the number of observations in $\{i : i \notin D_b^*\}$. The cross-validation estimators tend to slightly overestimate the prediction error, especially when the sample size of the training datasets is smaller than the sample size n (Gerds and Schumacher, 2007; Yuan, 2008). Some other estimators aim to retain the advantages of both the apparent loss method and cross-validation methods, such as the 0.632+ bootstrap estimator (Gerds and Schumacher, 2007).

The model-based estimate of the prediction error is defined by

$$\widehat{\mathrm{PE}}(t_0) = \frac{1}{n} \sum_{i=1}^{n} E_Y \left\{ \left(Y_i - \widehat{Y}_i(X_i; \widehat{\theta}) \right)^2 \right\} . \tag{3.7}$$

When the prediction model is misspecified, this model-based estimator may be seriously biased (Lawless and Yuan, 2010).

3.1.3 Estimating Prediction Error for Censored Data

Inverse Probability of Censoring Weights

While the prediction $\widehat{Y}_i$ can be always obtained from a prediction model, the response Y may be unknown due to censoring (i.e. when $C_i < t_0$). Either weighting or imputation is used to deal with this situation. Inverse probability of censoring weights (IPCW) can be useful to estimate the prediction error and the IPCW estimator of the apparent loss (3.4) is

$$\widehat{\mathrm{PE}}(t_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}\left\{Y_i - \widehat{Y}_i(X_i;\widehat{\theta})\right\}^2 , \tag{3.8}$$

where $\Delta_i = I(C_i > \min(T_i, t_0))$ indicates Y_i is known, and $\pi_i = E(\Delta_i|T_i, X_i)$ is the conditional expectation of Δ_i given (T_i, X_i). Given (T_i, X_i) the only random quantity in Δ_i is C_i, so $\pi_i = P(C_i > \min(T_i, t_0)|T_i, X_i) = \mathcal{G}(\min(T_i, t_0)|T_i, X_i)$, where $\mathcal{G}(\cdot)$ is the survivor function of the right censoring time. The motivation for this IPCW approach is explained as follows:

$$\begin{aligned}
E_{T,C,X}\left[\widehat{\mathrm{PE}}(t_0)\right] &= E_{T,X}\left\{E_{C|T,X}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}\left\{Y_i - \widehat{Y}_i(X_i;\widehat{\theta})\right\}^2\right]\right\} \\
&= E_{T,X}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{E_{C|T,X}\left[\Delta_i\right]}{\pi_i}\left\{Y_i - \widehat{Y}_i(X_i;\widehat{\theta})\right\}^2\right\} \\
&= E_{T,X}\left\{\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \widehat{Y}_i(X_i;\widehat{\theta})\right\}^2\right\} .
\end{aligned}$$

In practise, $\mathcal{G}(c|T_i, X_i)$ and π_i are unknown; but a consistent estimate of (3.8) is obtained if a $\sqrt{n}$-consistent estimate of π_i is used. Such an estimate can be obtained by using a

regression modelling approach based on the Cox model (Cox, 1975) or additive models (Aalen, 1989), for example.

ROC Curve and Estimation of the AUC

Another approach to examine the performance of a classification for survival status is based on the receiver operating characteristic (ROC) curve. Consider a set of binary predictors of survival status at a fixed time t_0, $\widehat{Y}(X;\theta) = I(\mathcal{F}(t_0|X;\widehat{\theta}) > c)$, where $\mathcal{F}(\cdot)$ is the survivor function of the event time and $c \in (0,1)$. The true positive rate (TPR) and false positive rate (FPR) are defined as

$$\begin{aligned} \text{TPR}(c) &= P(\widehat{Y} = 1|Y = 1) = \frac{P(\mathcal{F}(t_0|X;\widehat{\theta}) > c, T > t_0)}{P(T > t_0)} , \\ \text{FPR}(c) &= P(\widehat{Y} = 1|Y = 0) = \frac{P(\mathcal{F}(t_0|X;\widehat{\theta}) > c, T \leq t_0)}{P(T \leq t_0)} . \end{aligned} \tag{3.9}$$

The ROC curve is obtained by plotting TPR(c) against FPR(c) for values of c increasing from 0 to 1. The best possible prediction method would yield a point in the upper left corner at coordinate (0,1) of the ROC space (representing 100% sensitivity and 100% specificity). While a point along a diagonal line (the so-called line of no-discrimination) corresponds to a prediction scheme no better than a random guess. The area under curve (AUC) is a summary measure of ROC curve, which is equal to the probability that a predictor will rank a randomly chosen positive instance higher than a randomly chosen negative one. These probabilities can be estimated using an inverse probability of censoring weighting approach as well. For example,

$$\widehat{P}(\mathcal{F}(t_0|X;\widehat{\theta}) > c, T > t_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i}{\widehat{\pi}_i} I(\mathcal{F}(t_0|X_i;\widehat{\theta}) > c, T_i > t_0) .$$

An alternative approach is to consider a C-statistic (concordance statistic) (Heagerty and Zheng, 2005; Uno et al., 2011), which is defined as

$$\widehat{\text{AUC}} = P(g(X_i;\widehat{\theta}) < g(X_j;\widehat{\theta}) \mid T_i > T_j) \,, \tag{3.10}$$

where $g(X)$ can be the linear predictor $X'\beta$.

3.2 Prediction for Interval-Censored Data

Here we propose methods for characterizing the predictive accuracy of a regression model when the outcome of interest is an interval-censored event time. We let $[L_i, R_i]$ denote the censoring interval known to contain the event time T_i for subject i, then the observed data is $D = \{(L_i, R_i, X_i), i = 1, \ldots, n\}$. When data are incomplete, either weighting or imputation are commonly used. The model-based and imputation-based estimators of the prediction error are defined as follows:

$$\widehat{\text{PE}}_{\text{Model-Based}}(t_0) = \frac{1}{n}\sum_{i=1}^{n} E\left[\left\{Y_i - \widehat{Y}_i(X_i;\widehat{\theta})\right\}^2 |X_i\right] \,, \tag{3.11}$$

$$\begin{aligned}\widehat{\text{PE}}_{\text{Imputed}}(t_0) = \frac{1}{n}\sum_{i=1}^{n} \Bigg\{ &\Delta_i \left\{Y_i - \widehat{Y}_i(X_i;\widehat{\theta})\right\}^2 \\ &+ (1-\Delta_i)\, E\left[\left\{Y_i - \widehat{Y}_i(X_i;\widehat{\theta})\right\}^2 |T_i \in [L_i, R_i], X_i\right]\Bigg\} \,. \end{aligned} \tag{3.12}$$

These two estimators are entirely based on the prediction model, so their performance depends on the correct response model specification and consistent parameter estimation. We also consider the use of weighting. For right-censored data, the observed data is a function of both the event process and the censoring process. For the interval-censored case, the observed data is also influenced by the inspection process. In the sections that follow, we consider a multistate framework for joint consideration of the event, inspection

and censoring (drop-out) processes, and discuss inverse weighted and augmented inverse weighted techniques. We begin with a discussion of the joint model for the response and observation process.

3.2.1 Notation and Formulation of Observation Process Models

We consider a single individual, let $0 = V_0 < V_1 < \cdots < V_n$ denote the times of assessments since disease onset, and let $N(u) = \sum_{r=1}^{\infty} I(V_r \leq u)$ count the number of assessments at time u. Let C be a random drop-out time and $C(u) = I(u \leq C)$ indicate whether this individual is in cohort or not at time u. We also let T be the event time, X be a $p \times 1$ fixed covariate vector, $\{X(s), 0 < s\}$ be a time-dependent covariate process, $\bar{X}(u) = \left\{X(a_1), \ldots, X(a_{N(u^-)})\right\}$ be the history of the observed value at time $u > 0$, and $\bar{W}(u) = \left\{W(a_1), \ldots, W(a_{N(u^-)})\right\}$ denote the recorded event status at the process assessment here where $W(u) = I(T < u)$. The complete history observed at time s is then $\mathcal{H}(s) = \left\{(dN(u), C(u)), 0 < u < s, X, \bar{X}(s), \bar{W}(s)\right\}$. Since the goal is to use genetic data to predict the development of PsA, it is inappropriate to control for time-varying markers in the causal pathway in the model for the response process. Here we adopt a simple hazard function of the form

$$\lim_{\Delta t \downarrow 0} \frac{P(T < t + \Delta t | T \geq t, X)}{\Delta t} = I(t \leq T) h(t|X)$$

for the response model. The intensities for the inspection and censoring processes are meant to provide good representations of the data and so conditioning on all available data is appropriate; we let

$$\lim_{\Delta t \downarrow 0} \frac{P(\Delta N(t) = 1 | \mathcal{H}(t))}{\Delta t} = C(t) \lambda(t | \mathcal{H}(t)) \ ,$$

$$\lim_{\Delta t \downarrow 0} \frac{P(C < t + \Delta t | \mathcal{H}(t))}{\Delta t} = C(t) \lambda^c(t | \mathcal{H}(t))$$

represent the inspection and censoring intensities.

We define a multistate process $\{Z(s), 0 < s\}$ with a state space

$$\mathfrak{S} = \left\{\mathbb{V}_0, \mathbb{V}_1, \mathbb{V}_2, \ldots, \mathbb{V}_1^E, \mathbb{V}_2^E, \ldots, \mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_1^E, \mathbb{C}_2^E, \ldots, \mathbb{E}\right\}$$

for joint consideration of the event, inspection and censoring processes; see Figure 3.1. Here $\mathbb{E}$ denotes the event, $\mathbb{V}_r$ denotes the state of having had the rth assessment without previously having experienced the event, $\mathbb{V}_r^E$ denotes the rth assessment occurring immediately after the event, $\mathbb{C}_r$ denotes the random drop-out after the $(r-1)$st assessment without the event, and $\mathbb{C}_r^E$ denotes the random drop-out after the $(r-1)$st assessment and after the event; note here the convention where we use a superscript E to denote the states after the process has been passed through $\mathbb{E}$.

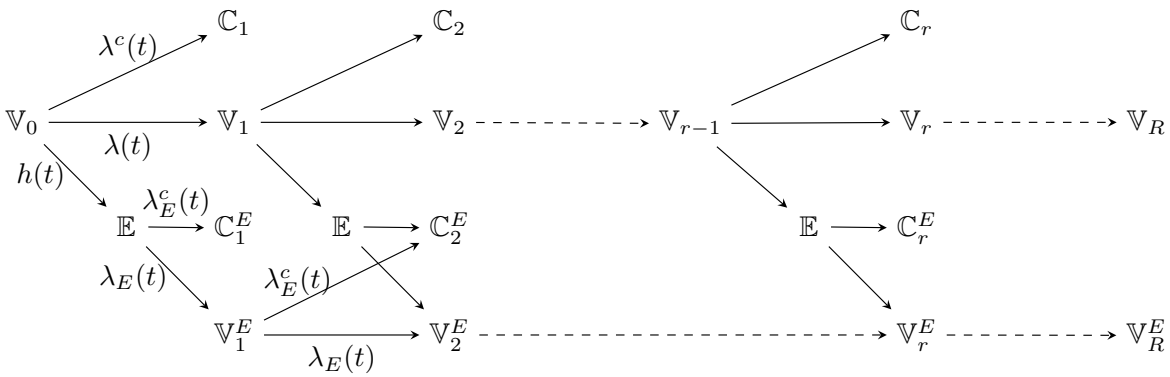


Figure 3.1: A multistate diagram for joint consideration of event, random drop-out and assessment times.

Following the occurrence of many transitions into non-absorbing states in $\mathfrak{S}$, the next transition to occur is governed by a competing risk process. The transition intensities are defined above with possible transition depicted by the arrows in Figure 3.1. We use a subscript E to denote an intensity post-event, that is, the intensity for an inspection

post-event is $\lambda_E(t)$ and the intensity for random drop-out post-event is $\lambda_E^c(t)$. If the event process is (conditionally) independent of the inspection process, then $\lambda_E(t) = \lambda(t)$, otherwise, we may assign a different intensity such as $\lambda_E(t) = \lambda(t)\exp(\alpha^s)$. Similarly, if the event process is independent of the censoring process, then $\lambda_E^c(t) = \lambda^c(t)$, otherwise, we may assign a different intensity such as $\lambda_E^c(t) = \lambda^c(t)\exp(\alpha^c)$.

3.2.2 Inverse Probability Weighted Estimator

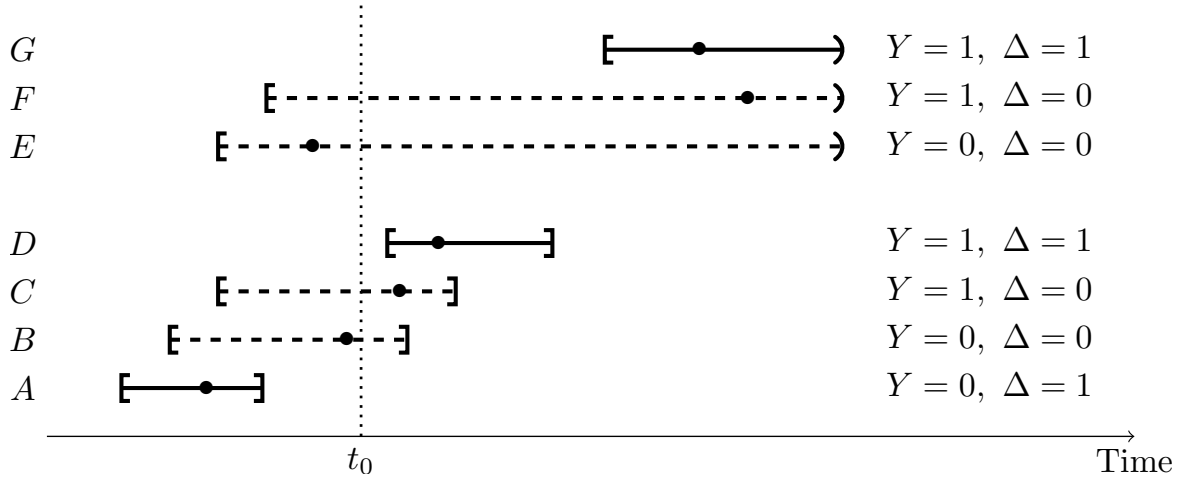


Figure 3.2: Four schematic diagram enumerating possible combinations of (Y, Δ); the solid lines denote observations in which the event status is known at t_0 and the dashed lines denote individual whose event status cannot be classified and hence who are excluded from the sum in (3.13); the solid dots denote the (unobserved) exact event times.

Figure 3.2 shows all the possible combinations of the event status and observation status indicators (Y, Δ). The IPW estimator of the prediction error is

$$\widehat{\mathrm{PE}}(t_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i}{\pi_i}\left\{Y_i - \widehat{Y}_i(X_i;\widehat{\theta})\right\}^2 , \tag{3.13}$$

where $\Delta_i = I(Y_i \text{ is known}) = I(t_0 \notin [L_i, R_i])$ and the weight is $\pi_i = E(\Delta_i|Y_i, X_i)$ which is the conditional expectation of Δ_i given (Y_i, X_i). Here the random variable Δ_i depends on the inspection process, the censoring process and the event process. The weight can then be written as

$$\pi_i = E(\Delta_i|Y_i, X_i) = E_{N,C,T|Y,X}[\Delta_i] = P(\Delta_i = 1|Y_i, X_i) \; . \tag{3.14}$$

The motivation for this IPW approach is explained by noting that

$$\begin{aligned} E_{N,C,T,Y,X}\left[\widehat{\mathrm{PE}}(t_0)\right] &= E_{Y,X}\left\{E_{N,C,T|Y,X}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}\left\{Y_i - \widehat{Y}_i(X_i;\widehat{\theta})\right\}^2\right]\right\} \\ &= E_{Y,X}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{E_{N,C,T|Y,X}\left[\Delta_i\right]}{\pi_i}\left\{Y_i - \widehat{Y}_i(X_i;\widehat{\theta})\right\}^2\right\} \\ &= E_{Y,X}\left\{\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \widehat{Y}_i(X_i;\widehat{\theta})\right\}^2\right\} \; . \end{aligned}$$

We explore this expectation in more detail separately for the case of $T \leq t_0$ and $T > t_0$.

Expectations Under the Condition $T \leq t_0$

After the occurrence of the event of interest (i.e. entry to an $\mathbb{E}$ state), the next event to occur can be a visit (i.e. entry to a $\mathbb{V}_r^E$ state) or censoring (i.e. entry to a $\mathbb{C}_r^E$ state). If Y is known, then a post-event assessment must be made before t_0. Thus $P(\Delta_i = 1|Y_i = 0, X_i) = P(\Delta_i = 1|T_i \leq t_0, X_i)$ can be written as

$$\begin{aligned} \int_0^{t_0}\left[\int_t^{t_0}\lambda_E(u|\mathcal{H}(u))\exp\left\{-\int_t^u \lambda_E(v|\mathcal{H}(v)) + \lambda_E^c(v|\mathcal{H}(v))\mathrm{d}v\right\}\mathrm{d}u\right] \\ \times f(t|T_i \leq t_0, X_i)\exp\left\{-\int_0^t \lambda^c(s|\mathcal{H}(s))\mathrm{d}s\right\}\,\mathrm{d}t \; . \end{aligned} \tag{3.15}$$

If we assume $\lambda_E(t) = \lambda(t)$ and $\lambda_E^c(t) = \lambda^c(t)$, then (3.15) becomes

$$\int_0^{t_0} \left[\int_t^{t_0} \lambda(u|\mathcal{H}(u)) \exp\left\{ -\int_t^u \lambda(v|\mathcal{H}(v))\ \mathrm{d}v - \int_0^u \lambda^c(v|\mathcal{H}(v))\mathrm{d}v \right\} \mathrm{d}u \right] \times f(t|T_i \leq t_0, X_i)\ \mathrm{d}t\ .$$

Expectations Under the Condition $T > t_0$

If Y is known to be one, in this case then there must be an assessment without disease after t_0, which can be represented by an entry to a $\mathbb{V}_r$ state, $r = 1, 2, \ldots$. Therefore $Z(t_0^-) = \mathbb{V}_{r-1}$ for some r and the next transition to occur can be a transition into states $\mathbb{V}_r$, $\mathbb{C}_r$ or $\mathbb{E}$. In this case, $P(\Delta_i = 1|Y_i = 1, X_i) = P(\Delta_i = 1|T_i > t_0, X_i)$ is

$$\int_{t_0}^{\infty} \lambda(u|\mathcal{H}(u)) \exp\left\{ -\int_{t_0}^u [\lambda(v|\mathcal{H}(v)) + h(v|Z)]\ \mathrm{d}v - \int_0^u \lambda^c(v|\mathcal{H}(v))\ \mathrm{d}v \right\}\ \mathrm{d}u\ . \quad (3.16)$$

3.2.3 Augmented Inverse Probability Weighted Estimator

The augmented inverse probability weighted (AIPW) estimator of the prediction error is defined as

$$\widehat{\mathrm{PE}}_{\mathrm{AIPW}}(t_0) = \frac{1}{n} \sum_{i=1}^n \left[\frac{\Delta_i}{\widehat{\pi}_i} \left\{ Y_i - \widehat{Y}_i(X_i; \widehat{\theta}) \right\}^2 + \left(1 - \frac{\Delta_i}{\widehat{\pi}_i} \right) \Psi(X_i) \right]\ ,$$

where $\Psi(X_i) = E\left[\left\{ Y_i - \widehat{Y}_i(X_i; \widehat{\theta}) \right\}^2 |X_i \right]$.

Usually when augmented inverse weighted estimators are defined a so-called "double-robustness" property is discussed which states that if the weight model or the response (prediction) model is correct then a consistent estimator for the parameter of interest is obtained. In the present setting the weight is dependent on the response model and

therefore if the response model is incorrect then the weight model must be incorrect. Therefore in the present setting the double robustness property is not present; see Section 5.3 for further comments. There is merit to investigating the empirical bias and relative efficiency of the estimator above however and we do so in what follows.

Properties of $\widehat{\mathbf{PE}}_{\mathbf{AIPW}}(t_0)$

If the weight is correctly modeled (i.e. we correctly jointly model (dN, C, T) on the test data), then

$$\begin{aligned}
E\left\{\widehat{\mathrm{PE}}_{\mathrm{AIPW}}\right\} =& E_{Y,X}\left[E_{N,C,T|Y,X}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\Delta_i}{\pi_i}\left\{Y_i-\widehat{Y}_i(X_i;\widehat{\theta})\right\}^2+\left(1-\frac{\Delta_i}{\pi_i}\right)\Psi(X_i)\right]\right\}\right] \\
=& E_{Y,X}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{E_{N,C,T|Y,X}(\Delta_i)}{\pi_i}\left\{Y_i-\widehat{Y}_i(X_i;\widehat{\theta})\right\}^2+\left(1-\frac{E_{N,C,T|Y,X}(\Delta_i)}{\pi_i}\right)\Psi(X_i)\right]\right] \\
=& E_{Y,X}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{P(\Delta_i=1|Y_i,X_i)}{\pi_i}\left\{Y_i-\widehat{Y}_i(X_i;\widehat{\theta})\right\}^2+\left(1-\frac{P(\Delta_i=1|Y_i,X_i)}{\pi_i}\right)\Psi(X_i)\right]\right] \\
=& E_{Y,X}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i-\widehat{Y}_i(X_i;\widehat{\theta})\right\}^2\right] .
\end{aligned}$$

However if only the prediction model is correctly modeled (i.e. only $f(t_i|x_i)$ is correctly modeled on training data), then

$$\begin{aligned}
E\left\{\widehat{\mathrm{PE}}_{\mathrm{AIPW}}\right\} =& E_{Y,X}\left[E_{N,C,T|Y,X}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\Delta_i}{\pi_i}\left\{Y_i-\widehat{Y}_i(X_i;\widehat{\theta})\right\}^2+\left(1-\frac{\Delta_i}{\pi_i}\right)\Psi(X_i)\right]\right\}\right] \\
=& E_{Y,X}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{E_{N,C,T|Y,X}(\Delta_i)}{\pi_i}\left\{Y_i-\widehat{Y}_i(X_i;\widehat{\theta})\right\}^2+\left(1-\frac{E_{N,C,T|Y,X}(\Delta_i)}{\pi_i}\right)\Psi(X_i)\right]\right] \\
=& E_{Y,X}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\pi_i^*}{\pi_i}\left\{Y_i-\widehat{Y}_i(X_i;\widehat{\theta})\right\}^2+\left(1-\frac{\pi_i^*}{\pi_i}\right)E_{Y|X}\left[\left\{Y_i-\widehat{Y}_i(X_i;\widehat{\theta})\right\}^2\right]\right]\right] \\
=& E_{Y,X}\left[\frac{1}{n}\sum_{i=1}^{n}E_{Y|X}\left[\left\{Y_i-\widehat{Y}_i(X_i;\widehat{\theta})\right\}^2\right]\right] \\
&+E_{Y,X}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\pi_i^*}{\pi_i}\left(\left\{Y_i-\widehat{Y}_i(X_i;\widehat{\theta})\right\}^2-E_{Y|X}\left[\left\{Y_i-\widehat{Y}_i(X_i;\widehat{\theta})\right\}^2\right]\right)\right] \\
=& E_{Y,X}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i-\widehat{Y}_i(X_i;\widehat{\theta})\right\}^2\right] .
\end{aligned}$$

The AIPW method therefore yields a consistent estimator of the prediction error under the usual condition that only the response model need be correct. It therefore represents an alternative approach to the direct model-based approach of (3.11) or the imputation approach of (3.12).

3.2.4 ROC Curves and the Area Under the Curve

The ROC curves and the AUC statistic can also be estimated by the IPW and AIPW approaches, as in the case of right-censoring. The weighting scheme and then the calculation of weights are the same as Section 3.2.2. Similarly, we can estimate those probabilities

separately using inverse weighting methods. For example,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\widehat{\pi}_i}I(\mathcal{F}(t_0|X_i;\widehat{\theta})>c, T_i>t_0)\ ,$$

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\Delta_i}{\widehat{\pi}_i}I(\mathcal{F}(t_0|X_i;\widehat{\theta})>c, T_i>t_0)+\left(1-\frac{\Delta_i}{\widehat{\pi}_i}\right)E_{Y|X}\left[I(\mathcal{F}(t_0|X_i;\widehat{\theta})>c, T_i>t_0)\right]\right\}$$

are the IPW and AIPW estimators of $P(\mathcal{F}(t_0|X;\widehat{\theta})>c, T>t_0)$ respectively.

3.3 Simulation Studies

3.3.1 Design and Results of Studies for Poisson Processes

We consider the setting with three covariates denoted X_{i1}, X_{i2} and X_{i3}. In one scenario they have marginal standard normal distributions with $X_{i1} \perp X_{i2}$, $X_{i1} \perp X_{i3}$, and $\text{corr}(X_{i2}, X_{i3}) = 0$ or 0.5. In a second scenario the covariates are binary with $P(X_{ij} = 1) = 0.5$, $j = 1, 2, 3$. The event time T_i follows a Weibull distribution given (X_{i1}, X_{i2}) with $\beta_1 = \log(2)$, $\beta_2 = \log(1.5)$ and shape $\kappa = 1.25$; that is,

$$\mathcal{F}(t|X_{i1}, X_{i2};\theta) = \exp\left\{-(\lambda t)^{\kappa}\exp\left(X_{i1}\beta_1 + X_{i2}\beta_2\right)\right\}\ ,$$

where $\theta = (\lambda, \kappa, \beta_1, \beta_2)'$; the value of λ is determined so that $P(T > 1) = 0.5$, where $P(T > 1) = E\{\mathcal{F}(t|X_{i1}, X_{i2};\theta)\}$. We consider an administrative censoring time τ such that $\mathcal{F}(\tau) = 0.9$. A time homogeneous Poisson process is used for the inspection process with rate

$$\lambda(s|X_{i1}, X_{i3};\gamma) = \exp(\gamma_0 + X_{i1}\gamma_1 + X_{i3}\gamma_2)\ ,$$

where $\gamma_1 = \log(1.1)$ and $\gamma_2 = \log(1.5)$ for the normal covariates and $\gamma_1 = \log(2)$ and $\gamma_2 = \log(2.5)$ for the binary covariates; γ_0 is determined to ensure that the average number of assessments by τ is controlled at $\mu = 10$ where $\mu = E\{\int_0^{\tau}\lambda(s|X_{i1}, X_{i3};\gamma)\mathrm{d}s\}$.

Let $0 = v_0 < v_1 < \ldots < v_n \leq \tau$ denote the inspection times, then the left and right endpoints of the censoring interval are $L = \max(v_r \cdot I(v_r < t))$ and $R = \min(v_r \cdot I(v_r > t))$ respectively. In the application there is no recorded right censoring time and so the expressions of (3.15) and (3.16) simplify to the following expressions (3.17) and (3.18)

$$\int_0^{t_0} \left[\int_t^{t_0} \lambda(u|\mathcal{H}(u)) \exp\left\{ - \int_t^u \lambda(v|\mathcal{H}(v)) \, \mathrm{d}v \right\} \mathrm{d}u \right] \times f(t|T_i \leq t_0, X_i) \, \mathrm{d}t \, , \tag{3.17}$$

$$\int_{t_0}^{\tau} \lambda(u|\mathcal{H}(u)) \exp\left[- \int_{t_0}^u \{\lambda(v|\mathcal{H}(v)) + h(v|X)\} \, \mathrm{d}v \right] \, \mathrm{d}u \, . \tag{3.18}$$

Thus the weights are estimated by modeling the event and inspection processes as described in the discussion of the simulation study. Datasets with sample sizes of $m = 500$ are simulated 100 times ($nsim = 100$) for each scenario. For each simulated dataset, parametric analyses were carried out to model the event time by using a Weibull distribution; both parametric and semiparametric analyses were used to model the gap times between two consecutive inspection times by an exponential distribution and Anderson-Gill model. The unweighted, model-based and imputed estimators are only based on the modeling of the event process; while the IPW and AIPW estimators depend on both the event and inspection processes, so the corresponding estimators are denoted as IPW-EXP, AIPW-EXP for parametric modeling and IPW-AG, AIPW-AG for semiparametric modeling. The empirical bias (EBIAS), the empirical standard error (ESE) and the relative empirical bias (%BIAS) of these estimators of the prediction error at time t_0 are summarized in Table 3.1 and Table 3.2, where t_0 values are taken to be the quartiles of the marginal distribution of T. The true prediction errors are estimated according to the formal definition evaluating

the expectation at the true parameter values. That is, we compute

$$\begin{aligned}\text{PE} &= E_{Y,X}\left[\{Y-\widehat{Y}(X;\theta)\}^2\right] \\ &= E_X\left\{E_{Y|X}\left[\{Y-\widehat{Y}(X;\theta)\}^2\right]\right\} \\ &= \int_x \left[\{1-\widehat{Y}(X;\theta)\}^2 P(T>t_0|X;\theta) + \{0-\widehat{Y}(X;\theta)\}^2 P(T\le t_0|X;\theta)\right] p(x)\mathrm{d}x\ .\end{aligned}$$

Under correct specification of the inspection model, the proposed IPW and AIPW estimators have relatively small biases compared to the unweighted estimators, while the variability (in terms of ESE) is greater. The AIPW estimators are more efficient than the IPW estimators; moreover none of the weighted estimators are as efficient as the model-based or imputation-based estimators. The misspecification of inspection model was next investigated by omitting one important covariate. In broad terms we found, as one would expect, that there was a consequent increase in the empirical bias of the IPW estimators, but that this bias remains smaller than that of the unweighted estimator for the misspecification considered here. Since the response model is correctly specified, the AIPW estimators under a misspecified inspection model have a comparable performance with those under correct specification of the inspection model; the bias remains small and the standard errors are very slightly lower in many cases than in the case when the inspection model is correctly specified.

3.3.2 Design and Results of Studies for Renewal Processes

Here we consider another scenario that the inspection process is governed by a non-Markov renewal process to further explore the influences of model misspecification of inspection process. The event times are generated as described in Section 3.3.1. The gap times between two consecutive inspections are generated by a Gamma distribution with shape

	Q_{25}				Q_{50}				Q_{75}			
METHOD	TRUE	EBIAS	ESE	%BIAS	TRUE	EBIAS	ESE	%BIAS	TRUE	EBIAS	ESE	%BIAS
					$X_{i2} \perp X_{i3}$							
Unweighted	0.2269	-0.0949	0.0190	-41.84	0.3063	-0.0425	0.0246	-13.86	0.2129	-0.0252	0.0178	-11.81
Model-Based	0.2269	-0.0005	0.0139	-0.21	0.3063	-0.0003	0.0120	-0.11	0.2129	0.0001	0.0105	0.07
IMPUTED	0.2269	-0.0003	0.0156	-0.15	0.3063	0.0006	0.0190	0.21	0.2129	0.0017	0.0164	0.79
					Correct[†1] Specification of Inspection Model							
IPW-EXP	0.2269	-0.0004	0.0332	-0.16	0.3063	0.0039	0.0286	1.28	0.2129	0.0008	0.0196	0.37
IPW-AG	0.2269	0.0000	0.0321	0.02	0.3063	0.0043	0.0282	1.39	0.2129	-0.0004	0.0193	-0.20
AIPW-EXP	0.2269	0.0006	0.0273	0.26	0.3063	0.0023	0.0251	0.76	0.2129	0.0015	0.0194	0.70
AIPW-AG	0.2269	0.0008	0.0268	0.35	0.3063	0.0027	0.0251	0.87	0.2129	0.0015	0.0193	0.71
					Misspecified[†2] Inspection Model							
IPW-EXP	0.2269	-0.0085	0.0321	-3.76	0.3063	-0.0098	0.0269	-3.20	0.2129	-0.0065	0.0191	-3.05
IPW-AG	0.2269	-0.0080	0.0309	-3.54	0.3063	-0.0095	0.0267	-3.09	0.2129	-0.0067	0.0189	-3.13
AIPW-EXP	0.2269	-0.0000	0.0263	-0.02	0.3063	0.0001	0.0243	0.05	0.2129	0.0013	0.0190	0.60
AIPW-AG	0.2269	0.0002	0.0257	0.08	0.3063	0.0003	0.0243	0.09	0.2129	0.0014	0.0189	0.68
					$X_{i2} \not\perp X_{i3}$							
Unweighted	0.2269	-0.0952	0.0172	-41.97	0.3063	-0.0448	0.0260	-14.63	0.2129	-0.0300	0.0172	-14.11
Model-Based	0.2269	-0.0039	0.0134	-1.70	0.3063	-0.0022	0.0137	-0.73	0.2129	-0.0000	0.0112	-0.02
IMPUTED	0.2269	-0.0044	0.0158	-1.94	0.3063	-0.0002	0.0202	-0.07	0.2129	-0.0004	0.0157	-0.21
					Correct[†1] Specification of Inspection Model							
IPW-EXP	0.2269	-0.0024	0.0301	-1.04	0.3063	0.0030	0.0297	0.97	0.2129	-0.0004	0.0197	-0.19
IPW-AG	0.2269	-0.0020	0.0288	-0.87	0.3063	0.0032	0.0296	1.06	0.2129	-0.0021	0.0192	-0.99
AIPW-EXP	0.2269	-0.0036	0.0257	-1.60	0.3063	0.0020	0.0260	0.65	0.2129	-0.0011	0.0192	-0.51
AIPW-AG	0.2269	-0.0036	0.0250	-1.61	0.3063	0.0024	0.0259	0.78	0.2129	-0.0009	0.0189	-0.42
					Misspecified[†2] Inspection Model							
IPW-EXP	0.2269	-0.0051	0.0272	-2.27	0.3063	-0.0105	0.0281	-3.44	0.2129	-0.0110	0.0186	-5.16
IPW-AG	0.2269	-0.0047	0.0261	-2.06	0.3063	-0.0103	0.0278	-3.36	0.2129	-0.0114	0.0183	-5.37
AIPW-EXP	0.2269	-0.0053	0.0237	-2.33	0.3063	-0.0014	0.0249	-0.46	0.2129	-0.0016	0.0184	-0.77
AIPW-AG	0.2269	-0.0053	0.0232	-2.33	0.3063	-0.0013	0.0248	-0.44	0.2129	-0.0016	0.0183	-0.75

* %BIAS reported are equal to EBIAS/TRUE$\times 10^2$

[†1] correct inspection model involves fitting $\lambda(s|X_{i1}, X_{i3}; \gamma) = \exp(\gamma_0 + X_{i1}\gamma_1 + X_{i3}\gamma_2)$,

[†2] misspecified inspection model involves fitting $\lambda(s|X_{i1}; \gamma) = \exp(\gamma_0 + X_{i1}\gamma_1)$.

Table 3.1: Empirical performance* of PE; sample size $m = 500$, number of simulations $nsim = 100$. The predictor is $\widehat{Y}(X; \widehat{\theta}) = I(P(T > t_0|X; \widehat{\theta}) > 0.5)$. The covariates have marginal standard normal distributions. The event time T_i follows a Weibull distribution with rate $\kappa\lambda(\lambda t)^{\kappa-1} \exp(X_{i1}\beta_1 + X_{i2}\beta_2)$, where $\beta_1 = \log(2)$, $\beta_2 = \log(1.5)$ and $\kappa = 1.25$. A time homogeneous Poisson process is used for the inspection process with rate $\exp(\gamma_0 + X_{i1}\gamma_1 + X_{i3}\gamma_2)$, where $\gamma_1 = \log(1.1)$ and $\gamma_2 = \log(1.5)$.

	Q_{25}				Q_{50}				Q_{75}			
METHOD	TRUE	EBIAS	ESE	%BIAS	TRUE	EBIAS	ESE	%BIAS	TRUE	EBIAS	ESE	%BIAS
					$X_{i2} \perp X_{i3}$							
Unweighted	0.2500	-0.0779	0.0196	-31.18	0.3836	-0.0358	0.0238	-9.32	0.2500	-0.0324	0.0205	-12.94
Model-Based	0.2500	0.0020	0.0180	0.78	0.3836	-0.0032	0.0178	-0.84	0.2500	-0.0016	0.0163	-0.65
IMPUTED	0.2500	0.0018	0.0184	0.71	0.3836	-0.0035	0.0201	-0.91	0.2500	-0.0010	0.0179	-0.42
					Correct[†1] Specification of Inspection Model							
IPW-EXP	0.2500	0.0030	0.0287	1.20	0.3836	-0.0014	0.0269	-0.37	0.2500	0.0009	0.0232	0.35
IPW-AG	0.2500	0.0015	0.0278	0.59	0.3836	-0.0013	0.0266	-0.34	0.2500	-0.0135	0.0214	-5.38
AIPW-EXP	0.2500	0.0013	0.0242	0.53	0.3836	-0.0033	0.0222	-0.86	0.2500	-0.0004	0.0206	-0.15
AIPW-AG	0.2500	0.0005	0.0234	0.21	0.3836	-0.0016	0.0222	-0.42	0.2500	-0.0091	0.0203	-3.64
					Misspecified[†2] Inspection Model							
IPW-EXP	0.2500	-0.0098	0.0248	-3.92	0.3836	-0.0188	0.0246	-4.89	0.2500	-0.0127	0.0214	-5.10
IPW-AG	0.2500	-0.0112	0.0242	-4.47	0.3836	-0.0186	0.0243	-4.84	0.2500	-0.0196	0.0205	-7.83
AIPW-EXP	0.2500	-0.0009	0.0224	-0.37	0.3836	-0.0033	0.0215	-0.87	0.2500	-0.0024	0.0199	-0.95
AIPW-AG	0.2500	-0.0019	0.0217	-0.76	0.3836	-0.0029	0.0215	-0.75	0.2500	-0.0065	0.0197	-2.61
					$X_{i2} \not\perp X_{i3}$							
Unweighted	0.2500	-0.0766	0.0167	-30.65	0.3836	-0.0340	0.0250	-8.87	0.2500	-0.0360	0.0203	-14.41
Model-Based	0.2500	-0.0002	0.0165	-0.09	0.3836	-0.0025	0.0149	-0.66	0.2500	-0.0018	0.0161	-0.72
IMPUTED	0.2500	-0.0007	0.0173	-0.26	0.3836	-0.0008	0.0201	-0.22	0.2500	-0.0023	0.0180	-0.91
					Correct[†1] Specification of Inspection Model							
IPW-EXP	0.2500	0.0033	0.0258	1.32	0.3836	-0.0012	0.0269	-0.32	0.2500	-0.0010	0.0217	-0.42
IPW-AG	0.2500	0.0024	0.0257	0.94	0.3836	-0.0010	0.0267	-0.27	0.2500	-0.0165	0.0203	-6.61
AIPW-EXP	0.2500	0.0010	0.0219	0.40	0.3836	-0.0021	0.0246	-0.55	0.2500	-0.0034	0.0208	-1.34
AIPW-AG	0.2500	0.0007	0.0220	0.26	0.3836	-0.0004	0.0245	-0.10	0.2500	-0.0124	0.0208	-4.96
					Misspecified[†2] Inspection Model							
IPW-EXP	0.2500	-0.0042	0.0217	-1.70	0.3836	-0.0164	0.0249	-4.28	0.2500	-0.0160	0.0205	-6.41
IPW-AG	0.2500	-0.0051	0.0217	-2.04	0.3836	-0.0161	0.0248	-4.19	0.2500	-0.0228	0.0198	-9.11
AIPW-EXP	0.2500	-0.0013	0.0188	-0.53	0.3836	-0.0020	0.0233	-0.52	0.2500	-0.0032	0.0208	-1.27
AIPW-AG	0.2500	-0.0019	0.0191	-0.77	0.3836	-0.0015	0.0233	-0.40	0.2500	-0.0073	0.0205	-2.93

* %BIAS reported are equal to EBIAS/TRUE$\times 10^2$

[†1] correct inspection model involves fitting $\lambda(s|X_{i1}, X_{i3}; \gamma) = \exp(\gamma_0 + X_{i1}\gamma_1 + X_{i3}\gamma_2)$,

[†2] misspecified inspection model involves fitting $\lambda(s|X_{i1}; \gamma) = \exp(\gamma_0 + X_{i1}\gamma_1)$.

Table 3.2: Empirical performance* of PE; sample size $m = 500$, number of simulations $nsim = 100$. The predictor is $\widehat{Y}(X; \widehat{\theta}) = I(P(T > t_0|X; \widehat{\theta}) > 0.5)$. The covariates are binary with $P(X_{ij} = 1) = 0.5$, $j = 1, 2, 3$. The event time T_i follows a Weibull distribution with rate $\kappa\lambda(\lambda t)^{\kappa-1} \exp(X_{i1}\beta_1 + X_{i2}\beta_2)$, where $\beta_1 = \log(2)$, $\beta_2 = \log(1.5)$ and $\kappa = 1.25$. A time homogeneous Poisson process is used for the inspection process with rate $\exp(\gamma_0 + X_{i1}\gamma_1 + X_{i3}\gamma_2)$, where $\gamma_1 = \log(2)$ and $\gamma_2 = \log(2.5)$.

η and rate $\exp(\gamma_0 + X_{i1}\gamma_1 + X_{i3}\gamma_2)$, where $\eta = 1.25$, 1.5 and 2 and the configurations of the other parameters are kept to be the same as in Section 3.3.1 for both the normal and binary covariates. For each combination of parameter configuration, 100 datasets with sample sizes equal to $m = 500$ are simulated.

Since the inspection process is non-Markov, the further the shape parameter η is from 1, the more different it is from a time homogeneous Poisson process and hence the greater the extent of misspecification. From Table 3.3, Table 3.4, Table 3.5 and Table 3.6, we found that empirical biases of the IPW-EXP and IPW-AG estimators increase as η increases, but the AIPW-EXP and AIPW-AG estimators maintain relatively small bias. The empirical standard errors of the AIPW estimators are smaller than those of the IPW estimators, which again demonstrates the improved efficiency of the AIPW estimators. As before none of the weighted or augmented weighted estimators perform as well as the model-based or imputation-based estimators of the prediction error since the correct response model is always used in the simulation studies.

We comment more on the possible utility of the augmented inverse probability weighted estimators in the settings where a validation sample is available in Chapter 5.

3.4 Application to the Psoriatic Arthritis Cohort

Our interest lies in identifying which among the 76 HLA markers are associated with increased risk of developing arthritis mutilans from the time of diagnosis of psoriatic arthritis and assessing the predictive performance of the models obtained by penalized regression through application of the methods in Section 3.2. While there is no clinical agreement on how to precisely define arthritis mutilans, it represents a state of significant joint damage arising from an extreme form of the disease; here we define it as present if an individual has

METHOD	Q_{25} TRUE	EBIAS	ESE	%BIAS	Q_{50} TRUE	EBIAS	ESE	%BIAS	Q_{75} TRUE	EBIAS	ESE	%BIAS
					$X_{i2} \perp X_{i3}$							
					Shape $\eta = 1.25$							
Unweighted	0.2269	-0.0973	0.0198	-42.90	0.3063	-0.0409	0.0216	-13.35	0.2129	-0.0267	0.0203	-12.56
Model-Based	0.2269	0.0000	0.0143	0.02	0.3063	-0.0008	0.0115	-0.27	0.2129	-0.0008	0.0120	-0.37
IMPUTED	0.2269	-0.0011	0.0178	-0.51	0.3063	0.0008	0.0172	0.27	0.2129	-0.0020	0.0189	-0.95
IPW-EXP	0.2269	0.0009	0.0349	0.40	0.3063	0.0143	0.0264	4.66	0.2129	0.0062	0.0223	2.93
IPW-AG	0.2269	0.0035	0.0350	1.55	0.3063	0.0135	0.0263	4.40	0.2129	0.0044	0.0219	2.09
AIPW-EXP	0.2269	-0.0039	0.0286	-1.73	0.3063	0.0044	0.0233	1.44	0.2129	-0.0013	0.0212	-0.59
AIPW-AG	0.2269	-0.0019	0.0287	-0.84	0.3063	0.0046	0.0232	1.49	0.2129	-0.0012	0.0210	-0.59
					Shape $\eta = 1.5$							
Unweighted	0.2269	-0.0933	0.0222	-41.13	0.3063	-0.0430	0.0229	-14.04	0.2129	-0.0231	0.0221	-10.86
Model-Based	0.2269	-0.0009	0.0149	-0.38	0.3063	-0.0019	0.0111	-0.60	0.2129	0.0012	0.0120	0.56
IMPUTED	0.2269	-0.0013	0.0199	-0.57	0.3063	-0.0013	0.0181	-0.42	0.2129	0.0015	0.0201	0.73
IPW-EXP	0.2269	0.0118	0.0364	5.19	0.3063	0.0182	0.0299	5.95	0.2129	0.0134	0.0260	6.30
IPW-AG	0.2269	0.0162	0.0365	7.16	0.3063	0.0167	0.0288	5.44	0.2129	0.0105	0.0258	4.91
AIPW-EXP	0.2269	0.0007	0.0309	0.31	0.3063	0.0015	0.0251	0.49	0.2129	0.0031	0.0243	1.45
AIPW-AG	0.2269	0.0039	0.0310	1.73	0.3063	0.0013	0.0248	0.43	0.2129	0.0025	0.0242	1.19
					Shape $\eta = 2$							
Unweighted	0.2269	-0.0948	0.0219	-41.78	0.3063	-0.0385	0.0209	-12.58	0.2129	-0.0210	0.0174	-9.88
Model-Based	0.2269	0.0001	0.0141	0.04	0.3063	0.0006	0.0111	0.18	0.2129	0.0007	0.0106	0.31
IMPUTED	0.2269	-0.0008	0.0182	-0.37	0.3063	0.0004	0.0168	0.14	0.2129	0.0025	0.0163	1.19
IPW-EXP	0.2269	0.0099	0.0366	4.37	0.3063	0.0333	0.0267	10.86	0.2129	0.0201	0.0204	9.44
IPW-AG	0.2269	0.0187	0.0362	8.24	0.3063	0.0304	0.0265	9.94	0.2129	0.0167	0.0201	7.85
AIPW-EXP	0.2269	-0.0056	0.0316	-2.47	0.3063	0.0046	0.0237	1.50	0.2129	0.0050	0.0190	2.33
AIPW-AG	0.2269	0.0008	0.0316	0.37	0.3063	0.0043	0.0236	1.40	0.2129	0.0045	0.0188	2.10

* %BIAS reported are equal to EBIAS/TRUE$\times 10^2$

Table 3.3: Empirical performance* of PE; sample size $m = 500$, number of simulations $nsim = 100$. The inspection process is generated by a renewal process. The gap times between two consecutive inspections are generated by a Gamma distribution with shape η and rate $\exp(\gamma_0 + X_{i1}\gamma_1 + X_{i3}\gamma_2)$, where $\gamma_1 = \log(1.1)$, $\gamma_2 = \log(1.5)$ and $\eta = 1.25$, 1.5 and 2. (Normal Cases, $X_{i2} \perp X_{i3}$)

METHOD	Q_{25} TRUE	EBIAS	ESE	%BIAS	Q_{50} TRUE	EBIAS	ESE	%BIAS	Q_{75} TRUE	EBIAS	ESE	%BIAS
					$X_{i2} \not\perp X_{i3}$							
					Shape $\eta = 1.25$							
Unweighted	0.2269	-0.0927	0.0195	-40.87	0.3063	-0.0483	0.0200	-15.78	0.2129	-0.0270	0.0200	-12.67
Model-Based	0.2269	-0.0017	0.0143	-0.76	0.3063	-0.0041	0.0120	-1.34	0.2129	-0.0015	0.0110	-0.72
IMPUTED	0.2269	-0.0021	0.0169	-0.92	0.3063	-0.0033	0.0168	-1.06	0.2129	0.0006	0.0180	0.30
IPW-EXP	0.2269	0.0046	0.0331	2.01	0.3063	0.0074	0.0229	2.42	0.2129	0.0091	0.0230	4.30
IPW-AG	0.2269	0.0085	0.0323	3.75	0.3063	0.0069	0.0228	2.24	0.2129	0.0065	0.0228	3.08
AIPW-EXP	0.2269	-0.0017	0.0285	-0.73	0.3063	-0.0030	0.0217	-0.97	0.2129	0.0023	0.0215	1.08
AIPW-AG	0.2269	0.0010	0.0278	0.45	0.3063	-0.0027	0.0217	-0.87	0.2129	0.0022	0.0213	1.04
					Shape $\eta = 1.5$							
Unweighted	0.2269	-0.0892	0.0208	-39.33	0.3063	-0.0408	0.0236	-13.33	0.2129	-0.0265	0.0164	-12.44
Model-Based	0.2269	0.0005	0.0153	0.23	0.3063	0.0011	0.0128	0.35	0.2129	-0.0010	0.0111	-0.47
IMPUTED	0.2269	0.0001	0.0188	0.04	0.3063	0.0026	0.0209	0.84	0.2129	0.0006	0.0152	0.26
IPW-EXP	0.2269	0.0161	0.0353	7.09	0.3063	0.0233	0.0260	7.61	0.2129	0.0130	0.0190	6.10
IPW-AG	0.2269	0.0202	0.0353	8.89	0.3063	0.0217	0.0260	7.08	0.2129	0.0104	0.0188	4.86
AIPW-EXP	0.2269	0.0031	0.0311	1.38	0.3063	0.0070	0.0250	2.27	0.2129	0.0021	0.0183	1.00
AIPW-AG	0.2269	0.0062	0.0310	2.74	0.3063	0.0070	0.0249	2.27	0.2129	0.0022	0.0181	1.05
					Shape $\eta = 2$							
Unweighted	0.2269	-0.0951	0.0185	-41.91	0.3063	-0.0428	0.0239	-13.98	0.2129	-0.0242	0.0192	-11.38
Model-Based	0.2269	-0.0025	0.0139	-1.11	0.3063	-0.0024	0.0131	-0.79	0.2129	-0.0008	0.0128	-0.39
IMPUTED	0.2269	-0.0043	0.0166	-1.88	0.3063	-0.0018	0.0202	-0.60	0.2129	0.0011	0.0191	0.51
IPW-EXP	0.2269	0.0074	0.0317	3.27	0.3063	0.0329	0.0292	10.73	0.2129	0.0209	0.0234	9.82
IPW-AG	0.2269	0.0159	0.0301	7.02	0.3063	0.0302	0.0289	9.87	0.2129	0.0171	0.0229	8.04
AIPW-EXP	0.2269	-0.0109	0.0264	-4.82	0.3063	0.0042	0.0261	1.36	0.2129	0.0046	0.0228	2.16
AIPW-AG	0.2269	-0.0049	0.0250	-2.17	0.3063	0.0040	0.0259	1.29	0.2129	0.0043	0.0224	2.03

* %BIAS reported are equal to EBIAS/TRUE$\times 10^2$

Table 3.4: Empirical performance* of PE; sample size $m = 500$, number of simulations $nsim = 100$. The inspection process is generated by a renewal process. The gap times between two consecutive inspections are generated by a Gamma distribution with shape η and rate $\exp(\gamma_0 + X_{i1}\gamma_1 + X_{i3}\gamma_2)$, where $\gamma_1 = \log(1.1)$, $\gamma_2 = \log(1.5)$ and $\eta = 1.25$, 1.5 and 2. (Normal Cases, $X_{i2} \not\perp X_{i3}$)

METHOD	Q_{25} TRUE	EBIAS	ESE	%BIAS	Q_{50} TRUE	EBIAS	ESE	%BIAS	Q_{75} TRUE	EBIAS	ESE	%BIAS
					$X_{i2} \perp X_{i3}$							
					Shape $\eta = 1.25$							
Unweighted	0.2500	-0.0772	0.0215	-30.90	0.3836	-0.0282	0.0246	-7.34	0.2500	-0.0346	0.0207	-13.82
Model-Based	0.2500	0.0016	0.0201	0.63	0.3836	-0.0003	0.0153	-0.08	0.2500	-0.0034	0.0170	-1.36
IMPUTED	0.2500	0.0008	0.0204	0.33	0.3836	0.0031	0.0196	0.82	0.2500	-0.0045	0.0191	-1.81
IPW-EXP	0.2500	0.0132	0.0295	5.27	0.3836	0.0190	0.0303	4.97	0.2500	0.0044	0.0227	1.77
IPW-AG	0.2500	0.0133	0.0300	5.33	0.3836	0.0177	0.0297	4.61	0.2500	-0.0108	0.0212	-4.33
AIPW-EXP	0.2500	0.0038	0.0249	1.52	0.3836	0.0047	0.0244	1.23	0.2500	-0.0051	0.0203	-2.05
AIPW-AG	0.2500	0.0046	0.0255	1.84	0.3836	0.0063	0.0242	1.63	0.2500	-0.0141	0.0200	-5.63
					Shape $\eta = 1.5$							
Unweighted	0.2500	-0.0761	0.0215	-30.45	0.3836	-0.0307	0.0247	-7.99	0.2500	-0.0263	0.0197	-10.51
Model-Based	0.2500	0.0022	0.0185	0.86	0.3836	-0.0003	0.0145	-0.09	0.2500	0.0006	0.0181	0.24
IMPUTED	0.2500	0.0023	0.0201	0.90	0.3836	0.0003	0.0218	0.07	0.2500	0.0008	0.0195	0.34
IPW-EXP	0.2500	0.0162	0.0323	6.46	0.3836	0.0237	0.0258	6.18	0.2500	0.0194	0.0216	7.76
IPW-AG	0.2500	0.0175	0.0319	6.99	0.3836	0.0219	0.0258	5.71	0.2500	0.0012	0.0194	0.46
AIPW-EXP	0.2500	0.0031	0.0267	1.25	0.3836	0.0004	0.0230	0.10	0.2500	0.0043	0.0196	1.71
AIPW-AG	0.2500	0.0048	0.0263	1.91	0.3836	0.0022	0.0231	0.57	0.2500	-0.0062	0.0184	-2.47
					Shape $\eta = 2$							
Unweighted	0.2500	-0.0774	0.0186	-30.94	0.3836	-0.0306	0.0274	-7.99	0.2500	-0.0269	0.0222	-10.74
Model-Based	0.2500	0.0006	0.0168	0.24	0.3836	-0.0004	0.0158	-0.10	0.2500	-0.0012	0.0171	-0.47
IMPUTED	0.2500	0.0000	0.0183	0.01	0.3836	0.0006	0.0224	0.15	0.2500	-0.0011	0.0211	-0.44
IPW-EXP	0.2500	0.0172	0.0323	6.88	0.3836	0.0356	0.0318	9.27	0.2500	0.0260	0.0249	10.40
IPW-AG	0.2500	0.0194	0.0320	7.75	0.3836	0.0324	0.0315	8.46	0.2500	0.0065	0.0236	2.60
AIPW-EXP	0.2500	-0.0007	0.0266	-0.29	0.3836	0.0007	0.0277	0.18	0.2500	0.0030	0.0226	1.22
AIPW-AG	0.2500	0.0023	0.0265	0.92	0.3836	0.0025	0.0276	0.65	0.2500	-0.0078	0.0223	-3.12

* %BIAS reported are equal to EBIAS/TRUE$\times 10^2$

Table 3.5: Empirical performance* of PE; sample size $m = 500$, number of simulations $nsim = 100$. The inspection process is generated by a renewal process. The gap times between two consecutive inspections are generated by a Gamma distribution with shape η and rate $\exp(\gamma_0 + X_{i1}\gamma_1 + X_{i3}\gamma_2)$, where $\gamma_1 = \log(2)$, $\gamma_2 = \log(2.5)$ and $\eta = 1.25$, 1.5 and 2. (Binary Cases, $X_{i2} \perp X_{i3}$)

METHOD	Q_{25} TRUE	EBIAS	ESE	%BIAS	Q_{50} TRUE	EBIAS	ESE	%BIAS	Q_{75} TRUE	EBIAS	ESE	%BIAS
					$X_{i2} \not\perp X_{i3}$							
					Shape $\eta = 1.25$							
Unweighted	0.2500	-0.0771	0.0185	-30.83	0.3836	-0.0358	0.0250	-9.34	0.2500	-0.0295	0.0218	-11.80
Model-Based	0.2500	-0.0015	0.0155	-0.61	0.3836	-0.0022	0.0159	-0.58	0.2500	0.0014	0.0156	0.55
IMPUTED	0.2500	-0.0016	0.0170	-0.65	0.3836	-0.0039	0.0211	-1.02	0.2500	0.0021	0.0193	0.82
IPW-EXP	0.2500	0.0066	0.0292	2.63	0.3836	0.0072	0.0289	1.89	0.2500	0.0142	0.0240	5.70
IPW-AG	0.2500	0.0075	0.0281	3.01	0.3836	0.0069	0.0291	1.81	0.2500	-0.0040	0.0227	-1.62
AIPW-EXP	0.2500	-0.0010	0.0235	-0.40	0.3836	-0.0056	0.0233	-1.47	0.2500	0.0040	0.0220	1.60
AIPW-AG	0.2500	0.0001	0.0229	0.04	0.3836	-0.0036	0.0233	-0.94	0.2500	-0.0062	0.0220	-2.47
					Shape $\eta = 1.5$							
Unweighted	0.2500	-0.0768	0.0195	-30.72	0.3836	-0.0285	0.0253	-7.43	0.2500	-0.0302	0.0214	-12.09
Model-Based	0.2500	-0.0006	0.0174	-0.23	0.3836	0.0009	0.0153	0.23	0.2500	-0.0005	0.0180	-0.19
IMPUTED	0.2500	-0.0006	0.0188	-0.23	0.3836	0.0004	0.0199	0.11	0.2500	-0.0000	0.0193	-0.01
IPW-EXP	0.2500	0.0107	0.0281	4.29	0.3836	0.0257	0.0287	6.70	0.2500	0.0179	0.0240	7.16
IPW-AG	0.2500	0.0133	0.0272	5.31	0.3836	0.0244	0.0288	6.37	0.2500	-0.0023	0.0213	-0.94
AIPW-EXP	0.2500	0.0000	0.0246	0.01	0.3836	0.0030	0.0252	0.79	0.2500	0.0032	0.0212	1.27
AIPW-AG	0.2500	0.0029	0.0236	1.14	0.3836	0.0051	0.0252	1.32	0.2500	-0.0082	0.0199	-3.30
					Shape $\eta = 2$							
Unweighted	0.2500	-0.0757	0.0180	-30.26	0.3836	-0.0274	0.0225	-7.14	0.2500	-0.0279	0.0205	-11.16
Model-Based	0.2500	-0.0021	0.0161	-0.82	0.3836	0.0008	0.0150	0.22	0.2500	0.0022	0.0170	0.86
IMPUTED	0.2500	-0.0027	0.0175	-1.07	0.3836	0.0017	0.0184	0.45	0.2500	0.0013	0.0197	0.51
IPW-EXP	0.2500	0.0232	0.0280	9.27	0.3836	0.0424	0.0300	11.06	0.2500	0.0280	0.0225	11.22
IPW-AG	0.2500	0.0247	0.0274	9.89	0.3836	0.0390	0.0288	10.17	0.2500	0.0057	0.0205	2.28
AIPW-EXP	0.2500	0.0032	0.0222	1.27	0.3836	0.0062	0.0237	1.62	0.2500	0.0043	0.0209	1.70
AIPW-AG	0.2500	0.0058	0.0220	2.33	0.3836	0.0080	0.0233	2.08	0.2500	-0.0079	0.0203	-3.15

* %BIAS reported are equal to EBIAS/TRUE$\times 10^2$

Table 3.6: Empirical performance* of PE; sample size $m = 500$, number of simulations $nsim = 100$. The inspection process is generated by a renewal process. The gap times between two consecutive inspections are generated by a Gamma distribution with shape η and rate $\exp(\gamma_0 + X_{i1}\gamma_1 + X_{i3}\gamma_2)$, where $\gamma_1 = \log(2)$, $\gamma_2 = \log(2.5)$ and $\eta = 1.25$, 1.5 and 2. (Binary Cases, $X_{i2} \not\perp X_{i3}$)

5 or more joints with the advanced stage of damage according to the modified Steinbrocker score. We consider data from 604 patients in the University of Toronto Psoriatic Arthritis Clinic with the median time from the diagnosis of PsA to last assessment being 12.5 years (lower quartile = 5.1, upper quartile = 21.5). Ninety-seven patients were known to develop arthritis mutilans because they had a visit with a damaged joint count of 5 or greater. The $25th$, $50th$ and $75th$ percentiles of the censoring interval lengths for these individuals were 2.50, 8.06 and 15.00 years respectively. We adopt a proportional hazards model with a piecewise constant (5-piece) baseline hazard with cut points at years 6.5, 10.5, 18 and 22. All models controlled for 6 clinical predictors including age at clinic entry, sex, age at onset of psoriasis, age at onset of PsA, family history of psoriasis and family history of psoriatic arthritis. The findings of all HLA variables selected by any method are listed in Table 2.3 of Chapter 2, the sign of the coefficients are consistent in the various final models.

We next apply the inverse weighting approach to estimate the prediction errors and the discriminative abilities for all the models. Semiparametric analysis was carried out to model the inspection process, that is, modeling the gap times between two consecutive inspection times by Anderson-Gill model. The covariates in the inspection process model are sex, age and the set of the 8 HLA markers. Figure 3.3 shows the results in terms of prediction error curves by using the unweighted, IPW and AIPW approaches (upper panels) and model- and imputation-based approaches (bottom panels), where time t_0, ranging from 0 to 40 years after the diagnosis of psoriatic arthritis. It is obvious that the unweighted estimates are greater than the weighted estimates since the unweighted estimators do not account for the unclassified portion in the sample. The estimates of the area under the ROC curves against time $t_0 = 5, 10, \ldots, 60$ by using unweighted, IPW and AIPW approaches are shown in Figure 3.4. The ROC curves at time $t_0 = 10$, 20 and 30 years after diagnosis of psoriatic arthritis for all three models with IPW (upper three panels) and AIPW (bottom three

panels) methods are shown in Figure 3.5; the summary statistic AUC is also given in the legend. We can conclude from these figures that the model by using ALASSO penalty has a better predictive performance in terms of a higher AUC compared with the models using LASSO and SCAD penalty functions.

3.5 Discussion and Future Research

In this chapter, we extend the methods of inverse probability weighting used for right-censored data to deal with interval-censored data arising from intermittent inspection of individuals. The simulation studies demonstrated that the proposed IPW and AIPW estimators led to better performance compared to simple methods of using the unweighted estimators. Note that many datasets involving interval-censored data report only the left and right endpoints of the censoring intervals; such information are insufficient to model the inspection process and implement the proposed method. Implementation hinges critically on the availability of all inspection times over the course of observation. In the data from the motivating study such data are available, and the proposed methods were illustrated by an application in which penalized regression was used to select the prediction model using methods of Wu and Cook (2015). The weighting methods were then applied to compare the performance of each model chosen using the different penalty functions.

Wu and Cook (2016) develop methods for variable selection with truncated and interval-censored data. While we have dealt with the latter complication here, it is less clear how one might assess predictive accuracy when samples are chosen subject to truncation, but this feature is often present in problems involving large datasets. We therefore plan to consider this in future research.

The availability of external validation data is also crucial when we consider the assess-

ment of the predictive performance. When there are no validation data, we can consider use of our proposed method as a tuning criterion for model/variable selection. While when the validation data are available, our proposed method can be used to examine the utility of the prediction model, which can be considered for use in adapting medical therapy based on patients individual risk. A common aim is to assess predictive accuracy using external validation samples. There are several clinical registries of individuals with psoriatic arthritis in Spain (Queiro et al., 2003), Ireland (Winchester et al., 2012), and Newfoundland (Rahman et al., 2011), most of which are devoted to some form of genetic research aiming to identify prognostic markers. We may consider use of these three external validation datasets and are currently investigating the extent of follow-up in these registries. An issue with the Newfoundland cohort is that the distribution of genetic markers and other attributes among the members of this registry is different than those individuals in the Toronto registry. As a result we might expect the estimated prediction error based on such an external validation sample to be quite different than the estimates obtained by cross-validation based on the Toronto registry.

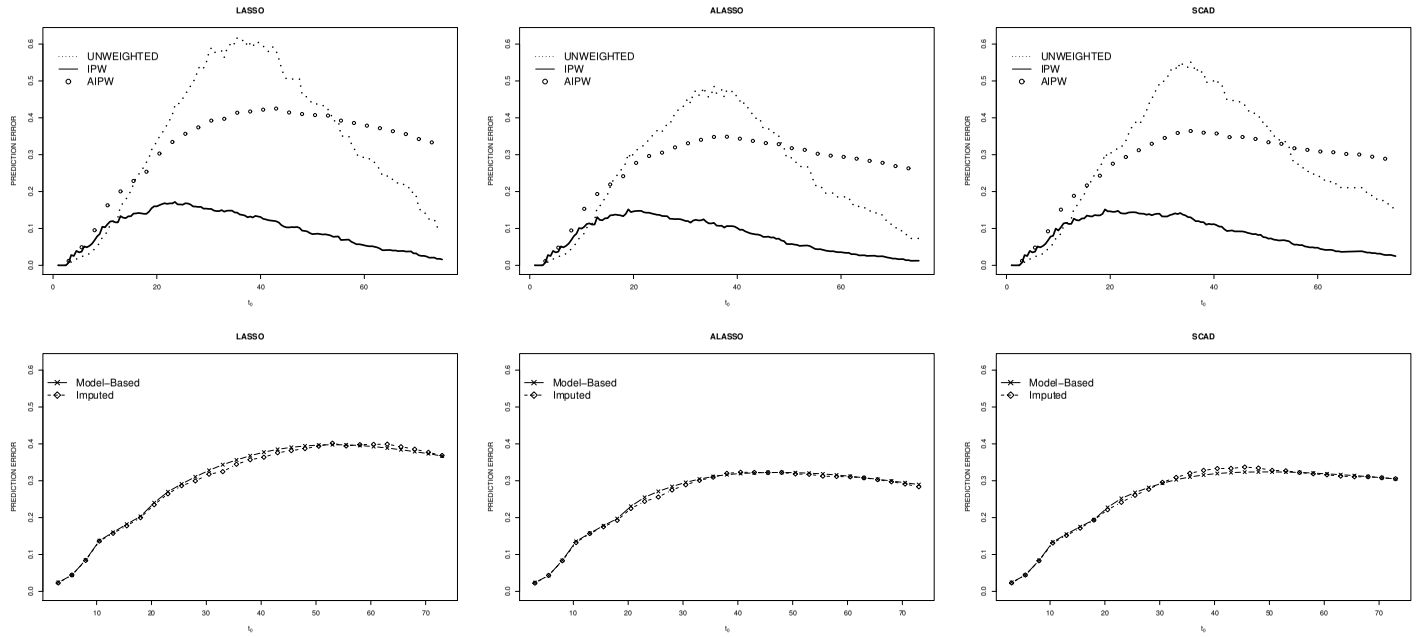


Figure 3.3: Plots of the estimates of the prediction error against t_0 with a binary predictor $I(P(T > t_0) > 0.5)$ for the models obtained from penalized regression with the LASSO, ALASSO and SCAD penalty functions. The upper panels show estimates obtained by using unweighted, inverse probability weighted and augmented inverse weighted methods; the bottom panels show estimates obtained by using model- and imputation-based methods.

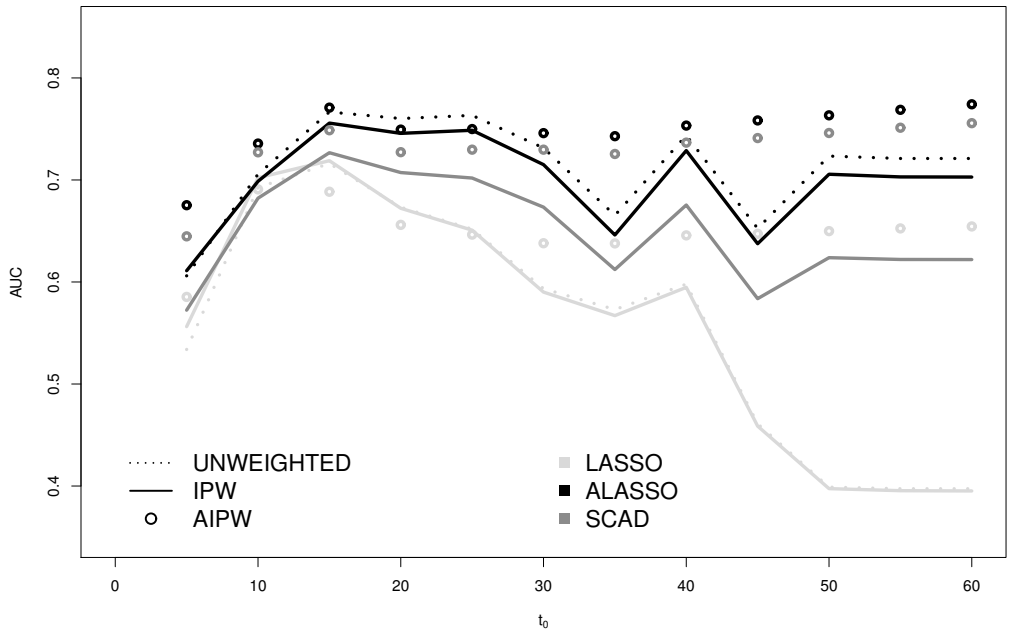


Figure 3.4: Plots of the estimates of the AUC against t_0 with a binary predictor $I(P(T > t_0) > c)$, where c ranging from 0 to 1. The response models are obtained from penalized regression with the LASSO, ALASSO and SCAD penalty functions.

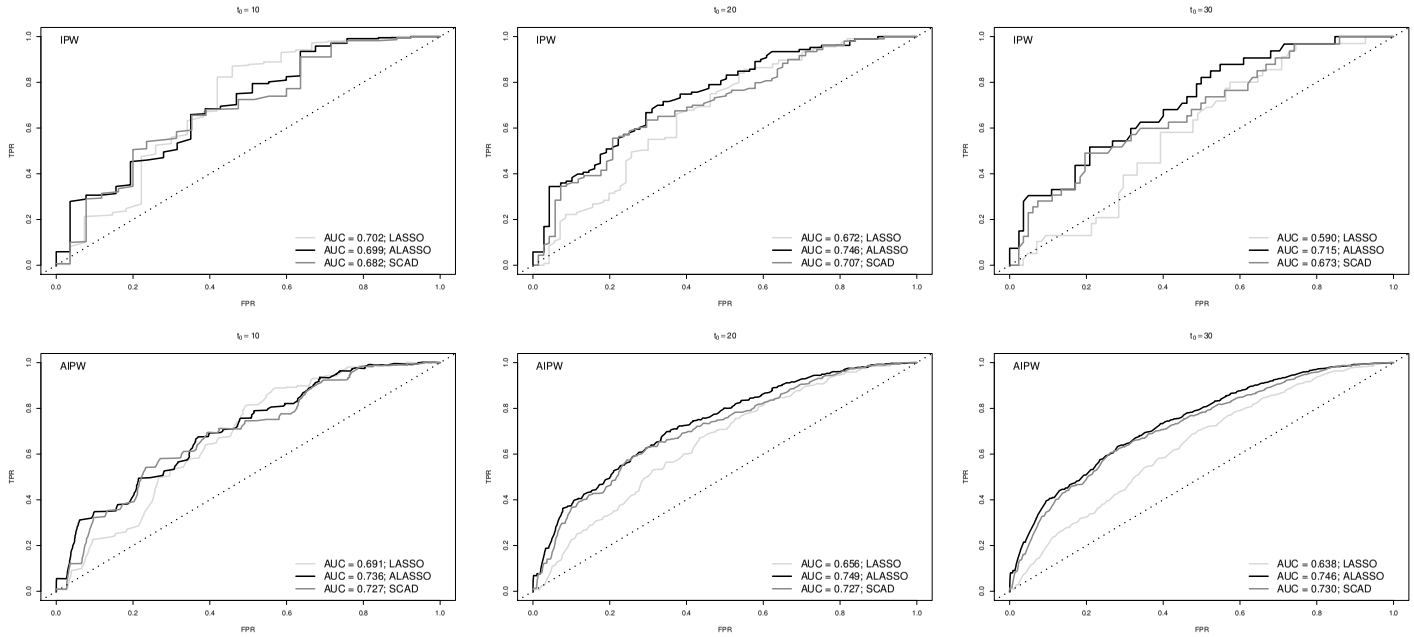


Figure 3.5: Estimates of ROC curves at t_0 =10, 20 and 30 years by inverse probability weighted (upper panels) and augmented inverse probability (bottom panels) methods for the models obtained from penalized regression with the LASSO, ALASSO and SCAD penalties with the tuning parameter selected by the 5-fold CV.

Chapter 4

A Two-Phase Model for Chronic Disease Processes Under Intermittent Inspection

4.1 Introduction

4.1.1 Disease Processes with Delayed Activity

Many chronic disease processes feature considerable variability in their course which must be dealt with in statistical analysis for valid inference. Regression modeling and regression diagnostics play a central role in explaining this variation in such a way that scientific understanding can be advanced. Another avenue is to generalize the family of stochastic models considered as the basis for analysis. Finite mixture models, for example, offer an appealing generalization as they involve conceptualization of two or more subpopula-

tions of individuals, each with different stochastic models generating the response process. When models are directed at dynamic aspects of disease processes the simplest and perhaps most studied mixture model accommodates a non-susceptible sub-population of individuals whose status will never change, while individuals in the complementary sub-population experience the disease process. Such models are often called cure-rate models when modeling the time to an event (Farewell, 1986) or mover-stayer models when considering multistate disease processes (Goodman, 1961; Frydman, 1984).

In many contexts it is unnatural to envision diseased individuals as being indefinitely at zero risk of disease progression. An alternative, and less extreme assumption is to consider two phases of the disease course: an inactive phase I during which diseased individuals do not experience clinically meaningful disease activity or damage and an active phase II of disease progression. Chronic diseases whose course can be represented in this way include HIV/AIDS where phase I represents the phase of HIV infection prior to the experience of AIDS defining events, and phase II represents the onset of opportunistic infections or death. In diabetes there may be a long phase I period during which no symptoms are evident, followed by a second phase during which there is evidence of retinopathy, nephropathy or other circulatory impairment. Individuals with hepatitic C infection may go a long time without experiencing any liver cirrhosis but will ultimately experience progressive liver damage. Finally, arthritis patients may simply have elevated markers of inflammation for some time before there is any evidence of joint damage, but once joint damage begins the risk of continued damage is substantially greater.

Phase I ends upon the occurrence of a precipitating event which signals the beginning of a fundamentally different phase (phase II) in which activity and damage are realized. The length of the phase I period may vary extensively between individuals and regression modeling techniques for time to event analysis can be adopted to explain this variation.

Once the period of morbidity begins, the nature of the morbidity process will drive the specification of the stochastic model for this phase. Often the dynamics of the disease process are sufficiently distinct in this phase that it is natural to define the time origin as the time of the transition from phase I to phase II. With this in mind we formulate a partially semi-Markov two-phase model in which one part is for the duration of phase I and another characterizes the dynamic disease process during phase II with the time origin being the start of the phase II. The term partially semi-Markov is because the time origin is only redefined once at the start of phase II. This model can be used to separately examine prognostic factors for the length of the inactive phase as well as factors prognostic for the nature and rate of change in the active phase. In some settings this will offer a more appropriate representation of complex multi-phase disease processes, can help identify different types of risk factors, and could yield more accurate prediction models.

The remainder of this chapter is organized as follows. In the next sub-section we describe the data from the University of Toronto Psoriatic Arthritis Cohort which motivates this work. In Section 4.2 we define notation and describe the two-phase model using a general multistate process to characterize the second phase where the time origin for the second phase is the time of the precipitative event. We also discuss likelihood construction when individuals are examined intermittently rendering the time of the precipitating event and subsequent transition interval-censored. In Section 4.3 we consider a special case of the general phase II model of Section 4.2 which is specified to correspond to the data from the motivating study. Specifically, the response of interest is intermittent counts of the number of damaged joint experienced by patients with a rheumatological disease, so we consider an analysis based on proportional rate models. We then develop an expectation-maximization algorithm (Dempster et al., 1977) for estimation under a model with piecewise constant intensities. A computationally more convenient two-stage estimation procedure is discussed

in Section 4.4 in which the parameters in the hazard for the end of phase I are estimated using standard likelihood for interval-censored data. The results of simulation studies examining the finite sample performance of estimators obtained by maximum likelihood and the two-stage procedure are given in Section 4.5, along with an application to the motivating study. Concluding remarks and topics for further research are provided in Section 4.6.

4.1.2 The University of Toronto Psoriatic Arthritis Cohort

The Centre for Prognosis Studies in Rheumatic Disease is a tertiary care center at the Toronto Western Hospital which treats patients with a variety of rheumatological conditions and maintains several clinic registries with prospective follow-up. One registry is of patients with psoriatic arthritis (PsA), an immunological disease which features both skin (psoriasis) and joint (arthritis) involvement. The psoriatic aspect of the condition arises from an overproduction of new skin cells resulting in red and white scaly patches of skin frequently located on the elbows, knees and scalp. As with other arthritic conditions, this disease can result in considerable inflammation and ultimately destruction of joints, which can lead to serious disability and poor quality of life (Chandran et al., 2010). This registry was established in 1976 and has been recruiting and following patients since its inception, and today it is one of the largest cohorts of patients with PsA in the world.

Patients in this registry undergo a detailed clinical and radiological examination upon entry to the clinic, and provide serum samples for genetic testing. Follow-up clinical and radiological assessments are scheduled annually and biannually respectively in order to track changes in joint damage. At each radiological assessment the degree of damage is recorded in sixty-four joints on a five-point scale (Rahman et al., 1998). To date 1191

patients have been recruited to the University of Toronto Psoriatic Arthritis Clinic. Of these 604 have undergone genetic testing to determine their human leukocyte antigen profile. Among these individuals the median time from clinic entry to the last radiological assessment is 6.3 years with a median of 3 radiological assessments per patient.

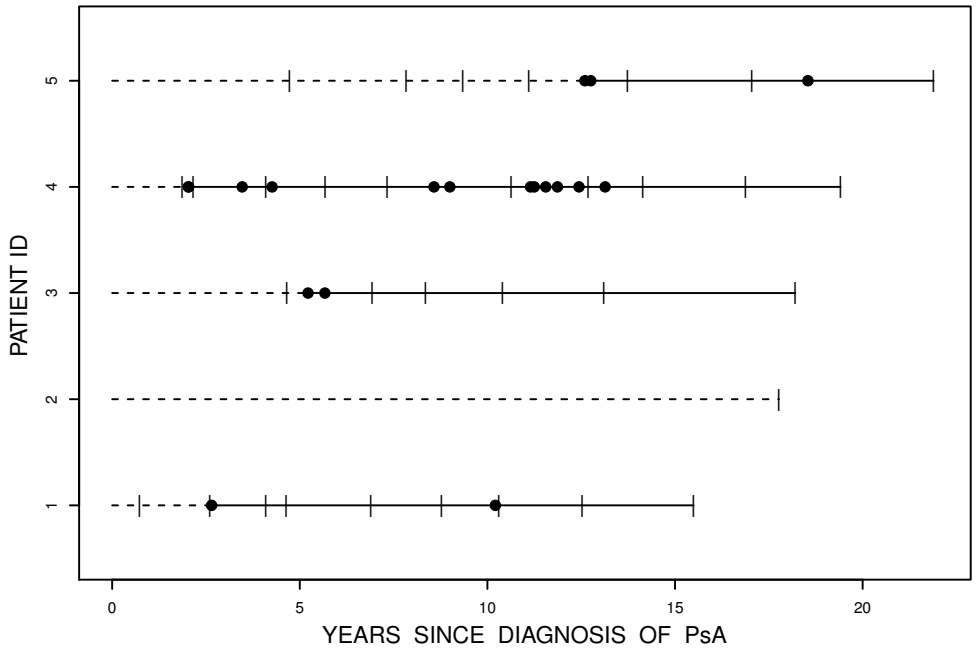


Figure 4.1: Plot of assessment times (hatch marks) and time of radiological damaged joints detected between assessments (solid points) from onset of psoriatic arthritis for a selected sample of patients from University of Toronto Psoriatic Arthritis Clinic. The dashed line denotes time from disease onset to first occurrence of joint damage, and the solid line denotes the period of disease progression following onset of damage.

We focus our modeling here on the accumulation of joint damage reflected by the total number of joints with at least grade 4 damage according to the Steinbrocker scoring

system (Wu and Cook, 2015). Figure 4.1 shows the time course of damage for a sample of five individuals. The horizontal axis is the time from disease onset and the length of the individual lines reflects the extent of follow-up of each individual; visits at which joint counts are made are represented by vertical tick marks. The dashed portion of each line reflects the period in which no joint damage is manifest and the solid lines reflect time following the occurrence of the first damaged joint. The precise times the joints became damaged are not available so for graphical illustration times were assigned by uniformly distributing them over the intervals during which they were known to occur; the dots are located at the resulting times. It is apparent that there are some individuals who experience active disease shortly after diagnosis (e.g. individuals 1 and 4) and some who enjoy a long period of time without damage (e.g. individuals 2 and 5). Moreover, once a patient develops their first damaged joint, some rapidly develop damage in other joints and for some individuals the rate of subsequent damage is very slow. These types of variation in manifestation of disease are what we accommodate with the two-phase model we describe in the next section.

4.2 Model Formulation and Likelihood under Intermittent Observation

4.2.1 General Formulation of a Two-Phase Model

We consider chronic diseases that feature a variable and potentially long phase I during which there are no clinically important manifestations of disease in affected individuals. If t denotes the time since disease onset, we let T_1 be a random variable representing the duration of phase I with t_1 representing its realization. During phase II where $t_1 < t$, in

general disease activity, it is evident through the occurrence of exacerbations or flares of symptoms, disability, or in the motivating context, joint damage and destruction. The variable duration of phase I and the distinct nature of the activity in the second phase suggests the use of $t^* = t - t_1$ as the time scale for the process in phase II, which is the time since the end of phase I. We let $\{Z(t^*), 0 < t^*\}$ be a multistate process with state space $\{1, 2, \ldots, \}$ reflecting the stage of the disease process in phase II. In Section 4.1 we consider the special case of a progressive multistate model which can alternatively be viewed as a recurrent event process.

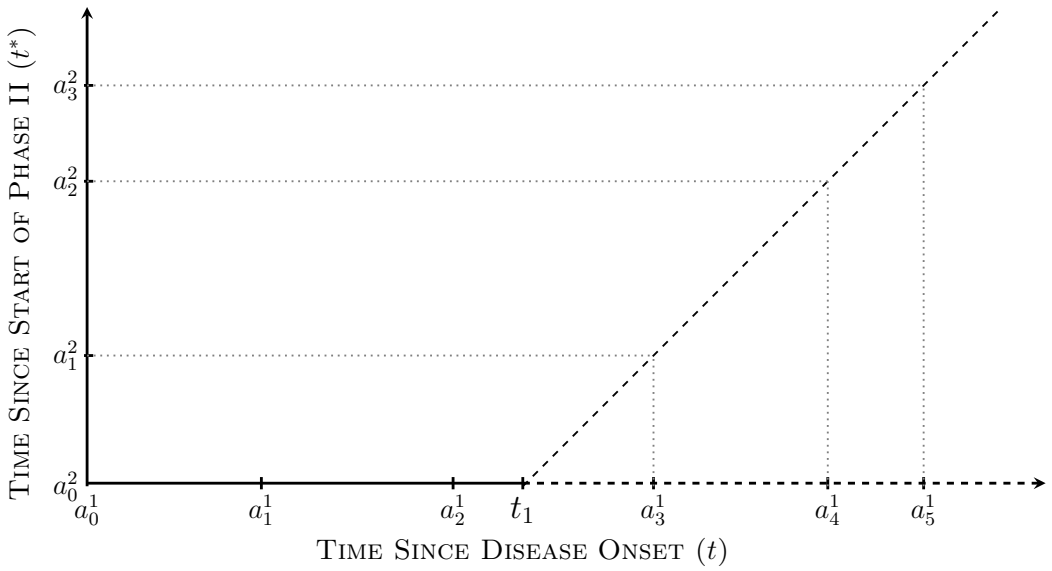


Figure 4.2: Lexis diagram of event and assessment times on the scale of disease duration (t) on the horizontal axis and the time since start of phase II (t^*) on the vertical axis.

To unify the notation for the two phases we augment the state space for the process in phase II to include a state 0 representing the status prior to T_1, and write $\bar{Z}(t) = Z(t^*)I(t_1 \leq t)$ and consider models for the stochastic process $\left\{\bar{Z}(t), 0 < t\right\}$; note that $Z(t^*) = \bar{Z}(t_1 + t^*)$. We let X be a $p \times 1$ vector of fixed covariates. The two-phase model

can be defined by first considering the hazard for the end of phase I, defined by

$$\lim_{\Delta t \downarrow 0} \frac{P(t \leq T_1 < t + \Delta t | t \leq T_1, X)}{\Delta t} = h(t|X) \; . \tag{4.1}$$

Covariate effects can be modeled using proportional (Cox, 1972), additive (Aalen, 1989), or hybrid Cox-Aalen models (Martinussen and Scheike, 2007).

We let $\mathcal{H}(t^*) = \{Z(u), 0 < u < t^*, x\}$ be the history of the process in phase II, and dynamic aspects of the process can be modeled through intensity functions (Andersen et al., 2012) given by

$$\lim_{\Delta t \downarrow 0} \frac{P\left(Z\left(t^* + \Delta t^-\right) = k | Z(t^{*-}) = j, \mathcal{H}(t^*)\right)}{\Delta t} = Y_j(t^*) \lambda_{jk}(t^* | \mathcal{H}(t^*)) \; , \tag{4.2}$$

where $Y_j(t^*) = I(Z(t^{*-}) = j)$, $j \in \{1, 2, \ldots\}$. If $\bar{\mathcal{H}}(t) = \left\{\bar{Z}(u), 0 < u < t, x\right\}$ denotes the history since the time of disease onset, then the intensity

$$\lim_{\Delta t \downarrow 0} \frac{P\left(\bar{Z}(t + \Delta t^-) = k | \bar{Z}(t^-) = j, \bar{\mathcal{H}}(t)\right)}{\Delta t} = \bar{Y}_j(t) \bar{\lambda}_{jk}(t | \bar{\mathcal{H}}(t)) \; , \tag{4.3}$$

governs the full process from disease onset, where $\bar{Y}_j(t) = I(\bar{Z}(t^-) = j)$ indicates whether an individual is at risk of a transition out of state $j \in \{0, 1, \ldots\}$ at time t. Note that if we denote (4.1) as $\lambda_{01}(t|\bar{\mathcal{H}}(t)\}$, then we can write $\bar{\lambda}_{jk}(t|\bar{\mathcal{H}}(t)) = \lambda_{jk}(B(t)|\bar{\mathcal{H}}(t))$ where $B(t) = I(t \leq t_1)t + I(t_1 < t)(t - t_1)$. Thus the process $\{\bar{Z}(t), 0 < t\}$ has a countable number of states in the state space and a semi-Markov feature in that the relevant time scale for the second phase of the disease process is the time since the end of phase I. With this time scale the process in phase II is Markov, but we refer to the process as a whole as partially semi-Markov.

The probability of a particular path $\mathcal{P}$ of this multistate process given $X = x$ is

$$\prod_{j=0}^{\infty} \prod_{k \in \mathcal{Z}_j} \left[\left\{ \prod_{t_r \in \bar{\mathcal{D}}_{jk}} \bar{\lambda}_{jk}(t_r | \bar{\mathcal{H}}(t_r)) \right\} \exp\left(- \int_0^{\infty} \bar{Y}_j(u) \bar{\lambda}_{jk}(u | \bar{\mathcal{H}}(u)) \mathrm{d}u \right) \right] \; , \tag{4.4}$$

where $\mathcal{Z}_j$ is the set of states that can be entered directly from state j and $\bar{\mathcal{D}}_{jk}$ is the set of $j \to k$ transition times (Andersen and Keiding, 2002) . This can be written more explicitly as

$$\begin{aligned} &\lambda_{01}(t_1|x) \exp\left(-\int_0^\infty \bar{Y}_0(u)\lambda_{01}(u|x)\mathrm{d}u\right) \\ &\times \left[\prod_{j=1}^{\infty}\prod_{k\in\mathcal{Z}_k}\left\{\prod_{t_r^*\in\mathcal{D}_{jk}}\lambda_{jk}(t_r^*|\mathcal{H}(t_r^*))\right\}\exp\left(-\int_0^\infty Y_j(u)\lambda_{jk}(u|\mathcal{H}(u))\mathrm{d}u\right)\right] , \end{aligned} \tag{4.5}$$

where if $t_r \in \bar{\mathcal{D}}_{jk}$, each $t_r^* \in \mathcal{D}_{jk}$ can be expressible as $t_r^* = t_r - t_1$. A slightly modified version of this probability expression can be derived for likelihood contributions when processes are under conditionally independent and non-informative censoring. Instead of pursuing this we consider next the problem of estimation and inference when such processes are under intermittent observation so that all event times are interval-censored.

4.2.2 Intermittent Assessment and Interval-Censored Data

Here we consider the case in which individuals are assessed intermittently and discuss the construction of the likelihood contribution for a single individual. Let $a_0 = 0$ denote the onset time of disease and $a_1 < \cdots < a_R$ denote the times of the R assessments at which point the individual's condition, and hence response status, is determined. The observed history at a_r^- is denoted by $H(a_r) = \left\{(a_\ell, \bar{Z}(a_\ell)), \ell = 0, 1, \ldots, r-1, X\right\}$, where we use a standard font for $H(\cdot)$ to distinguish it from the history of the process in continuous time. With fixed covariates, the full likelihood is

$$L \propto P\left(\bar{Z}(a_0), A_0 = a_0, X\right) \times \prod_{r=1}^{R} P\left(\bar{Z}(a_r), A_r = a_r|H(a_r)\right) . \tag{4.6}$$

We can omit the first term in the full likelihood if we condition on the covariate and the state occupied (0) at the onset of disease. We also assume the "sequential missing at

random" condition (Hogan et al., 2004) holds so that if an individual is observed up to a_{r-1}, then conditional on the event history at that time, the probability they are lost to follow-up and not observed at a_r cannot depend on events in $[a_{r-1}, a_r)$. We also assume the event process and inspection process are conditionally independent and that the inspection process is non-informative. Under these assumptions, we can focus on the partial likelihood of the form

$$L \propto \prod_{r=1}^{R} P\left(\bar{Z}(a_r)|H(a_r)\right) \ . \tag{4.7}$$

This observed data partial likelihood (4.7) can be maximized directly, but this can be challenging if the dimension of parameters is high and the expression of this likelihood is complicated due to intermittent assessment. Therefore, an expectation-maximization (EM) algorithm (Dempster et al., 1977) can alternatively be used with a complete data likelihood analogous to observed data likelihood where missing variables, in this case the transition time from phase I to phase II, are part of the complete data. This is a particularly attractive approach for the setting of piecewise constant intensities which we consider in the next section.

4.3 Piecewise Constant Baseline Functions and the EM Algorithm

4.3.1 Complete Data Log-Likelihood

The complete data likelihood (4.5) is given in general form for the case in which we consider the event times as observed, or subject at most to right-censoring. In this section, we redefine the notation by giving a superscript 1 or 2 to denote the phase. Here we consider

the setting with interval-censored data in phase II and let $a_0^1 = 0$ denote the onset of disease and $a_1^1 < \cdots < a_{R_1}^1$ denote the times of R_1 assessments at which point the individual's disease stage is determined. For information in phase II it is helpful to let $a_0^2 = 0$ denote the start time of phase II and $a_1^2 < \cdots < a_{R_2}^2$ denote the times of the R_2 radiological assessments during phase II. With the process in phase II a recurrent event process we can also let $n_r = \bar{Z}(a_r^2) - \bar{Z}(a_{r-1}^2)$ denote the number of events over the interval $\mathcal{A}_r = (a_{r-1}^2, a_r^2]$, $r = 1, \ldots, R_2$.

We adopt a Poisson process model for phase II such that $\lambda_{k,k+1}(t^*|\mathcal{H}(t^*)) = \rho(t^*|x)$ and write the complete data likelihood (4.5) as

$$\begin{aligned} L \propto \lambda_{01}(t_1|X) \exp\left(-\int_0^\infty \bar{Y}_0(u)\lambda_{01}(u|X)\mathrm{d}u\right) \\ \times \prod_{r=1}^{R_2}\left[\frac{1}{n_r!}\left\{\int_{a_{r-1}^2}^{a_r^2} \rho(u|X)\mathrm{d}u\right\}^{n_r} \exp\left\{-\int_{a_{r-1}^2}^{a_r^2} \rho(u|X)\mathrm{d}u\right\}\right] . \end{aligned} \tag{4.8}$$

We consider multiplicative models of the form $\lambda_{01}(t_1|X;\theta_1) = h_0(t_1;\alpha_1)\exp(X'\beta_1)$ for $T_1|X$ and $\lambda_{k,k+1}(t^*|\mathcal{H}(t^*)) = \rho_0(t^*;\alpha_2)\exp(X'\beta_2)$ for the recurrent event process in phase II, where $k \geq 1$, where α_1 indexes the baseline hazard function, α_2 indexes the baseline rate function, $\theta_1 = (\alpha_1', \beta_1')'$, $\theta_2 = (\alpha_2', \beta_2')'$ and $\theta = (\theta_1', \theta_2')'$. A weakly parametric piecewise exponential baseline hazard is adopted for the duration of phase I and a piecewise constant baseline rate model is adopted for the recurrent event process during phase II. These require specification of break-points where the baseline hazard and rate functions can take on different values and we denote these by $0 = b_0^1 < b_1^1 < \cdots < b_{K_1}^1$ and $0 = b_0^2 < b_1^2 < \cdots < b_{K_2}^2$ respectively. Then we let

$$\begin{aligned} h_0(t;\alpha_1) &= \alpha_{1k} \quad \text{if } t \in \mathcal{B}_k^1 = [b_{k-1}^1, b_k^1) \quad k = 1, \ldots, K_1, \\ \rho_0(t^*;\alpha_2) &= \alpha_{2k} \quad \text{if } t^* \in \mathcal{B}_k^2 = [b_{k-1}^2, b_k^2) \quad k = 1, \ldots, K_2 , \end{aligned} \tag{4.9}$$

respectively. We consider all the subintervals $\mathcal{C}_{rk} = \mathcal{A}_r \cap \mathcal{B}_k^2$ of length u_{rk} and let n_{rk} denote the unobserved number of events over $\mathcal{C}_{rk}$ such that $\sum_{k=1}^{K_2} n_{rk} = n_r \ r = 1, \ldots, R_2$. Since $N_{rk}|T_1, X \sim \text{Poisson}(\mu_{rk})$, where $\mu_{rk} = \alpha_{2k} u_{rk} \exp(X'\beta_2)$, then

$$E\left(N_{rk}|T_1, X, N_r\right) = n_r \cdot \alpha_{2k} u_{rk} / \sum_{k=1}^{K_2} \alpha_{2k} u_{rk} \ .$$

The complete data log likelihood is then

$$\log L_{\text{C}}(\theta) = \log L_{\text{C1}}(\theta_1) + \log L_{\text{C2}}(\theta_2) \ , \tag{4.10}$$

where

$$\log L_{\text{C1}}(\theta_1) = \delta_1 \left\{ \sum_{k=1}^{K_1} I_k(t_1) \left(\log \alpha_{1k} + X'\beta_1\right) - \sum_{k=1}^{K_1} \alpha_{1k} W_k(t_1) e^{X'\beta_1} \right\} - (1 - \delta_1) \sum_{k=1}^{K_1} \alpha_{1k} W_k(a_{R_1}) e^{X'\beta_1} \ , \tag{4.11}$$

$$\log L_{\text{C2}}(\theta_2) = \delta_1 \left\{ \sum_{r=1}^{R_2} \sum_{k=1}^{K_2} n_{rk} \left(\log \alpha_{2k} + X'\beta_2\right) - \sum_{r=1}^{R_2} \sum_{k=1}^{K_2} \alpha_{2k} u_{rk} e^{X'\beta_2} \right\} \ , \tag{4.12}$$

$I_k(u) = I(u \in \mathcal{B}_k^1)$ and $W_k(u) = \int_0^u I_k(s)\mathrm{d}s$ is the total time at risk in $\mathcal{B}_k^1$ over the interval $(0, u]$, $k = 1, \ldots, K_1$.

4.3.2 The EM Algorithm for Maximum Likelihood Estimation

At the vth iteration of the expectation-maximization (EM) algorithm, the E-Step is to take the conditional expectation

$$Q(\theta; \theta^{(v)}) = Q_1(\theta_1; \theta^{(v)}) + Q_2(\theta_2; \theta^{(v)}) \ , \tag{4.13}$$

where $Q_1(\theta_1; \theta^{(v)}) = E\left[\log L_{\text{C1}}(\theta_1)|D; \theta^{(v)}\right]$ and $Q_2(\theta_2; \theta^{(v)}) = E\left[\log L_{\text{C2}}(\theta_2)|D; \theta^{(v)}\right]$, where the observed data is $D = \left\{(a_r, \bar{Z}(a_r)), r = 0, 1, \ldots, R_1, X\right\}$. The unobserved quantities in

the complete data log likelihood $I_k(t_1)$, $W_k(t_1)$, n_{rk} and u_{rk} are all functions of T_1. Thus, their conditional expectations given the current estimates of parameters and the observed data D can be evaluated through

$$f_{t_1|D}(t_1|D;\theta) = \frac{f_1(t_1) \times \prod_{r=1}^{R_2} f_2(n_r|t_1)}{\int_{L_1}^{R_1} f_1(u_1) \times \prod_{r=1}^{R_2} f_2(n_r|u_1)\ \mathrm{d}u_1}\ , \tag{4.14}$$

where

$$f(t_1|X) = \prod_{k=1}^{K_1} \left\{ [\alpha_{1k} \exp(X'\beta_1)]^{I_k(t_1)} \cdot \exp\left(-\alpha_{1k} W_k(t_1) \exp(X'\beta_1)\right) \right\}$$

and

$$f_2(n_r|t_1, X) = \left[\sum_{k=1}^{K_2} \alpha_{2k} u_{rk} \exp(X'\beta_2) \right]^{n_r} \cdot \exp\left(- \sum_{k=1}^{K_2} \alpha_{2k} u_{rk} \exp(X'\beta_2) \right)\ .$$

The M-Step involves maximizing $Q(\theta;\theta^{(v)})$ with respect to θ and gets the updated estimate $\theta^{(v+1)}$. By reparametrization, we can write $Q(\theta;\theta^{(v)})$ in the form of a Poisson log-likelihood and use existing software for generalized linear model to maximize following the creation of a pseudo-dataset. We iterate between the E-step and M-step until the convergence criterion $\left|\left(\theta^{(v+1)} - \theta^{(v)}\right)/\theta^{(v)}\right| < \epsilon$ is achieved where ϵ is the user-specified tolerance. The details of the EM algorithm and the calculation of conditional expectations are given in Appendix 4.A.

4.3.3 Louis' Method for Estimates Obtained by Simultaneous Maximization

Here we describe how to implement Louis' (Louis, 1982) method based on the identity

$$I_{\mathrm{OBS}}(\theta) = \sum_{i=1}^{m} E[I_{\mathrm{C}i}(\theta)|D_i] - \sum_{i=1}^{m} E[S_i(\theta)S_i'(\theta)|D_i] + \sum_{i=1}^{m} E[S_i(\theta)|D_i]\{E[S_i(\theta)|D_i]\}'\ . \tag{4.15}$$

For simplicity, hereafter we drop the subscript i and only consider a single observation. Then the complete data score function, obtained from (4.10), is

$$S(\theta) = (S_1'(\theta_1), S_2'(\theta_2))' \ , \tag{4.16}$$

where $S_1(\theta_1) = (S_{11}'(\theta_1), S_{12}'(\theta_1))'$ with $S_{11}(\theta_1) = \partial \log L_{\mathrm{C}}(\theta)/\partial\alpha_1$ and $S_{12}(\theta_1) = \partial \log L_{\mathrm{C}}(\theta)/\partial\beta_1$, and $S_2(\theta_2) = (S_{21}'(\theta_2), S_{22}'(\theta_2))'$ with $S_{21}(\theta_2) = \partial \log L_{\mathrm{C}}(\theta)/\partial\alpha_2$ and $S_{22}(\theta_2) = \partial \log L_{\mathrm{C}}(\theta)/\partial\beta_2$. The corresponding contribution to the complete data information matrix is then

$$I_i = -\begin{pmatrix} \dfrac{\partial^2 \log L_{\mathrm{C}}(\theta)}{\partial\alpha_1\partial\alpha_1'} & \dfrac{\partial^2 \log L_{\mathrm{C}}(\theta)}{\partial\alpha_1\partial\beta_1'} & 0 & 0 \\ \dfrac{\partial^2 \log L_{\mathrm{C}}(\theta)}{\partial\beta_1\partial\alpha_1'} & \dfrac{\partial^2 \log L_{\mathrm{C}}(\theta)}{\partial\beta_1\partial\beta_1'} & 0 & 0 \\ 0 & 0 & \dfrac{\partial^2 \log L_{\mathrm{C}}(\theta)}{\partial\alpha_2\partial\alpha_2'} & \dfrac{\partial^2 \log L_{\mathrm{C}}(\theta)}{\partial\alpha_2\partial\beta_2'} \\ 0 & 0 & \dfrac{\partial^2 \log L_{\mathrm{C}}(\theta)}{\partial\beta_2\partial\alpha_2'} & \dfrac{\partial^2 \log L_{\mathrm{C}}(\theta)}{\partial\beta_2\partial\beta_2'} \end{pmatrix}.$$

For the specific models we consider,

$$\frac{\partial \log L_{\mathrm{C}}(\theta)}{\partial \alpha_{1k}} = \delta_1 \left\{ \frac{I_k(t_1)}{\alpha_{1k}} - \alpha_{1k} S_k(t_1) e^{x'\beta_1} \right\} - (1-\delta_1)\alpha_{1k} S_k(a_{R_1}) e^{x'\beta_1} \ , \quad k = 1, 2, \ldots, K_1 \ ,$$

$$\frac{\partial \log L_{\mathrm{C}}(\theta)}{\partial \beta_1} = \delta_1 \left\{ \sum_{k=1}^{K_1} \left(I_k(t_1) - \alpha_{1k} S_k(t_1) e^{x'\beta_1} \right) \right\} x - (1-\delta_1) \left\{ \sum_{k=1}^{K_1} \alpha_{1k} S_k(a_{R_1}) e^{x'\beta_1} \right\} x \ ,$$

$$\frac{\partial \log L_{\mathrm{C}}(\theta)}{\partial \alpha_{2k}} = \delta_1 \sum_{r=1}^{R_2} \left(\frac{n_{rk}}{\alpha_{2k}} - u_{rk} e^{x'\beta_2} \right) = \delta_1 \sum_{r=1}^{R_2} \left(\frac{n_r u_{rk}}{\sum_{j=1}^{K_2} \alpha_{2j} u_{rj}} - u_{rk} e^{x'\beta_2} \right) \ , \quad k = 1, 2, \ldots, K_2 \ ,$$

$$\frac{\partial \log L_{\mathrm{C}}(\theta)}{\partial \beta_2} = \delta_1 \sum_{r=1}^{R_2} \sum_{k=1}^{K_2} \left(n_{rk} - \alpha_{2k} u_{rk} e^{x'\beta_2} \right) x = \delta_1 \left(n - \sum_{r=1}^{R_2} \sum_{k=1}^{K_2} \alpha_{2k} u_{rk} e^{x'\beta_2} \right) x \ ,$$

$$- \frac{\partial^2 \log L_{\mathrm{C}}(\theta)}{\partial \alpha_{1j} \partial \alpha_{1k}} = 0 \ , \quad j \neq k \ ,$$

$$- \frac{\partial^2 \log L_{\mathrm{C}}(\theta)}{\partial \alpha_{1k}^2} = \delta_1 \frac{I_k(t_1)}{\alpha_{1k}^2} \ , \quad k = 1, 2, \ldots, K_1 \ ,$$

$$- \frac{\partial^2 \log L_{\mathrm{C}}(\theta)}{\partial \beta_1 \partial \alpha_{1k}} = \delta_1 S_k(t_1) e^{x'\beta_1} x + (1-\delta_1) S_k(a_{R_1}) e^{x'\beta_1} x \ ,$$

$$- \frac{\partial^2 \log L_{\mathrm{C}}(\theta)}{\partial \beta_1 \partial \beta_1'} = \delta_1 \left\{ \sum_{k=1}^{K_1} \alpha_{1k} S_k(t_1) e^{x'\beta_1} \right\} x x' + (1-\delta_1) \left\{ \sum_{k=1}^{K_1} \alpha_{1k} S_k(a_{R_1}) e^{x'\beta_1} \right\} x x' \ ,$$

$$- \frac{\partial^2 \log L_{\mathrm{C}}(\theta)}{\partial \alpha_{2k} \partial \alpha_{2\ell}} = \delta_1 \sum_{r=1}^{R_2} \frac{n_r u_{rk} u_{r\ell}}{\left(\sum_{j=1}^{K_2} \alpha_{2j} u_{rj} \right)^2} \ , \quad k, \ell = 1, 2, \ldots, K_2 \ ,$$

$$- \frac{\partial^2 \log L_{\mathrm{C}}(\theta)}{\partial \beta_2 \partial \alpha_{2k}} = \delta_1 \sum_{r=1}^{R_2} \sum_{k=1}^{K_2} u_{rk} e^{x'\beta_2} x \ ,$$

$$- \frac{\partial^2 \log L_{\mathrm{C}}(\theta)}{\partial \beta_2 \partial \beta_2'} = \delta_1 \sum_{r=1}^{R_2} \sum_{k=1}^{K_2} \alpha_{2k} u_{rk} e^{x'\beta_2} x x' \ .$$

The conditional expectations can be evaluated by Gaussian Quadrature described above and based on the conditional probability density function (4.14). To obtain the variance estimate, we need the inverse of the observed information matrix, thus in this case, we need to invert a high dimensional matrix. Instead of using the *solve* function, we used *ginv* function in *MASS* library (or *chol2inv* function).

4.4 Two-Stage Estimation

4.4.1 Description of Two-Stage Procedure

Instead of simultaneously estimating all the parameters in the full likelihood function (4.10), a two-stage estimation procedure can be adopted. Under this approach in the first stage we note that we can simply view T_1 as an interval-censored failure time with a hazard function indexed by θ_1. For this the pertinent data can be denoted by $\mathcal{C}_1 = [L_1, R_1)$, the interval known to contain T_1, and X. Here we let $Q_I(\cdot)$ denote the corresponding function in (4.13) under a two-stage procedure where

$$Q_{\text{I}}(\theta_1; \theta_1^{(v)}) = \delta_1 \left[\sum_{k=1}^{K_1} \hat{\iota}_k^{(v)} \left(\log \alpha_{1k} + X'\beta_1 \right) - \sum_{k=1}^{K_1} \widehat{\omega}_k^{(v)} \alpha_{1k} e^{X'\beta_1} \right] - (1 - \delta_1) \sum_{k=1}^{K_1} \alpha_{1k} W_k(a_{R_1}) e^{X'\beta_1} , \tag{4.17}$$

$\hat{\iota}_k^{(v)} = E\{I_k(t_1)|\mathcal{C}_1, x; \theta_1^{(v)}\}$ and $\widehat{\omega}_k^{(v)} = E\{W_k(t_1)|\mathcal{C}_1, x; \theta_1^{(v)}\}$. The conditional distribution of T_1 given $\mathcal{C}_1$ and X takes on a simpler form in this framework with $f(t_1|\mathcal{C}_1, X; \theta_1^{(v)}) = f_1(t_1) / \int_{L_1}^{R_1} f_1(u_1) du_1$ given by

$$f(t_1|\mathcal{C}_1, X; \theta_1^{(v)}) = \frac{[\prod_{k=1}^{K_1} \alpha_{1k}^{I_k(t_1)}] \times \exp(-\sum_{k=1}^{K_1} \alpha_{1k} W_k(t_1) \exp(X'\beta_1))}{\int_{L_1}^{R_1} [\prod_{k=1}^{K_1} \alpha_{1k}^{I_k(u_1)}] \times \exp(-\sum_{k=1}^{K_1} \alpha_{1k} W_k(u_1) \exp(X'\beta_1)) du_1} . \tag{4.18}$$

The expectations are therefore easier to carry out, and the maximization step is as before. Specifically (4.17) can be written as a Poisson log-likelihood and existing software can be used to maximize it following the creation of a pseudo-dataset.

In the second stage, θ_2 can be estimated via a modified expectation-maximization algorithm defined by plugging in the estimates of $\widehat{\theta}_1$ from stage one into the function $Q_{\text{II}}(\theta_2; \widehat{\theta}_1, \theta_2^{(v)})$ defined as

$$Q_{\text{II}}(\theta_2; \widehat{\theta}_1, \theta_2^{(v)}) = \sum_{r=1}^{R_2} \sum_{k=1}^{K_2} \delta_1 \left\{ \widehat{n}_{rk}^{(v)} \left(\log \alpha_{2k} + X'\beta_2 \right) - \alpha_{2k} \widehat{u}_{rk}^{(v)} \exp(X'\beta_2) \right\} , \tag{4.19}$$

where $\widehat{u}_{rk} = E[u_{rk}|D; \widehat{\theta}_1, \theta_2^{(v)}]$ and $\widehat{n}_{rk}^{(v)} = E[n_r \alpha_{2k}^{(v)} u_{rk} / \sum_{k=1}^{K_2} \alpha_{2k}^{(v)} u_{rk} | D; \widehat{\theta}_1, \theta_2^{(v)}]$. The conditional expectations in the second stage are based on (4.14) evaluated at $\widehat{\theta}_1$ and $\theta_2^{(v)}$ given the full set of observed data D. The objective function (4.19) can be rewritten to take the form of a Poisson log-likelihood and maximized using existing software as before. This two-stage estimation approach is quite similar to the method of simultaneous estimation we described in Section 4.3.2; however, this approach is computationally easier especially when the number of parameters is large. We comment further on the potential uses of this two-stage procedure in the Discussion.

4.4.2 Variance Estimation following Two-Stage Estimation

We estimate the asymptotic covariance matrix in the spirit of parametric two-stage estimation procedure (Newey and McFadden, 1994). The complete data score functions are shown in (4.16). For the simultaneous estimation approach, we solve the following estimating functions:

$$U(\theta) = 0 \ ,$$

where

$$U(\theta) = \begin{pmatrix} U_{11}(\theta_1, \theta_2) \\ U_{12}(\theta_1, \theta_2) \\ U_{21}(\theta_1, \theta_2) \\ U_{22}(\theta_1, \theta_2) \end{pmatrix} = \begin{pmatrix} E\left[S_{11}(\theta_1)|D; \theta_1, \theta_2\right] \\ E\left[S_{12}(\theta_1)|D; \theta_1, \theta_2\right] \\ E\left[S_{21}(\theta_2)|D; \theta_1, \theta_2\right] \\ E\left[S_{22}(\theta_2)|D; \theta_1, \theta_2\right] \end{pmatrix} \ . \tag{4.20}$$

For the two-stage estimation approach, in the first stage we solve

$$U_1^*(\theta_1) = \begin{pmatrix} U_{11}^*(\theta_1) \\ U_{12}^*(\theta_1) \end{pmatrix} = \begin{pmatrix} E\left[S_{11}(\theta_1)|\mathcal{C}_1, X; \theta_1\right] \\ E\left[S_{12}(\theta_1)|\mathcal{C}_1, X; \theta_1\right] \end{pmatrix} = 0 \ , \tag{4.21}$$

and in the second stage we solve

$$U_2^*(\theta_2) = \begin{pmatrix} U_{21}^*(\theta_2) \\ U_{22}^*(\theta_2) \end{pmatrix} = \begin{pmatrix} E[S_{21}(\theta_2)|D; \widehat{\theta}_1, \theta_2] \\ E[S_{22}(\theta_2)|D; \widehat{\theta}_1, \theta_2] \end{pmatrix} = 0 \; . \tag{4.22}$$

Thus, at the second stage of the two-stage procedure we plug $\widehat{\theta}_1$ into (4.20) and estimate θ_2 by solving the resulting equation. Let $U_1(\theta_1, \theta_2) = (U_{11}'(\theta_1, \theta_2), U_{12}'(\theta_1, \theta_2))'$, $U_2(\theta_1, \theta_2) = (U_{21}'(\theta_1, \theta_2), U_{22}'(\theta_1, \theta_2))'$, and $U_2^*(\theta_2) = U_2(\widehat{\theta}_1, \theta_2)$. Then if $\theta_0 = (\theta_{10}', \theta_{20}')'$ denotes the true value of θ, consider the Taylor expansion of the score function $U_2^*(\theta_2)$ around θ_0 and evaluate it at $\widehat{\theta}_2$ giving,

$$\begin{aligned} 0 = U_2^*(\widehat{\theta}_2) &= U_2(\widehat{\theta}_1, \widehat{\theta}_2) \\ &= U_2(\theta_{10}, \theta_{20}) + \frac{\partial U_2(\theta_{10}, \theta_{20})}{\partial \theta_1} \left(\widehat{\theta}_1 - \theta_{10}\right) + \frac{\partial U_2(\theta_{10}, \theta_{20})}{\partial \theta_2} \left(\widehat{\theta}_2 - \theta_{20}\right) + o_p(n^{1/2}) \; . \end{aligned}$$

Also

$$0 = U_1^*(\widehat{\theta}_1) = U_1^*(\theta_{10}) + \frac{\partial U_1^*(\theta_{10})}{\partial \theta_1} \left(\widehat{\theta}_1 - \theta_{10}\right) + o_p(n^{1/2}) \; ,$$

therefore,

$$\begin{pmatrix} U_1^*(\theta_{10}) \\ U_2(\theta_{10}, \theta_{20}) \end{pmatrix} = - \begin{pmatrix} \dfrac{\partial U_1^*(\theta_{10})}{\partial \theta_1} & 0 \\ \dfrac{\partial U_2(\theta_{10}, \theta_{20})}{\partial \theta_1} & \dfrac{\partial U_2(\theta_{10}, \theta_{20})}{\partial \theta_2} \end{pmatrix} \begin{pmatrix} \widehat{\theta}_1 - \theta_{10} \\ \widehat{\theta}_2 - \theta_{20} \end{pmatrix} \; .$$

As $n \to \infty$, by the law of large numbers

$$-\frac{1}{n} \begin{pmatrix} \dfrac{\partial U_1^*(\theta_{10})}{\partial \theta_1} & 0 \\ \dfrac{\partial U_2(\theta_{10}, \theta_{20})}{\partial \theta_1} & \dfrac{\partial U_2(\theta_{10}, \theta_{20})}{\partial \theta_2} \end{pmatrix} \xrightarrow{p} \begin{pmatrix} I_{11}^* & 0 \\ I_{21} & I_{22} \end{pmatrix} \triangleq A \; ,$$

and by the central limit theorem,

$$\frac{1}{\sqrt{n}} \begin{pmatrix} U_1^*(\theta_{10}) \\ U_2(\theta_{10}, \theta_{20}) \end{pmatrix} \xrightarrow{d} N(0, B) \; , \quad \text{where} \;\; B = \begin{pmatrix} I_{11}^* & 0 \\ 0 & I_{22} \end{pmatrix} \; .$$

Therefore,

$$\sqrt{n}\begin{pmatrix}\widehat{\theta}_1-\theta_{10}\\ \widehat{\theta}_2-\theta_{20}\end{pmatrix}\xrightarrow{d}N(0,A^{-1}BA^{-1'})\ .$$

4.5 Simulation Studies and Application

4.5.1 Design and Interpretation of Simulation Studies

In this section, a simulation study is conducted to demonstrate the performance of proposed two-phase model. For each individual i, a $p\times 1$ covariate vector X_i is generated from a multivariate normal distribution with mean 0 and a covariance matrix Σ, where $p=2$, $\Sigma_{ij}=\varrho^{|i-j|}$ and $\varrho=0.5$. The duration of the indolent phase T_1 is generated from an exponential distribution with rate $\alpha_1\exp(X_i'\beta_1)$. We set $\beta_1=(0.5,0.5)'$ and solve for the value of the baseline rate α_1 such that $F(C)=\int_x P(T_1<C|x)P(x)dx=0.8$, where $C=50$ is the administrative censoring time. The gap times between the consecutive events are generated by an exponential distribution with rate $\alpha_2\exp(X_i'\beta_2)$, where $\beta_2=(-0.5,-0.5)$ and $\alpha_2=0.5$.

We let R_i denote the number of assessments for individual i, which is generated according to a truncated Poisson distribution to ensure at least one follow-up assessment, with

$$P(R_i=r_i|R_i\geq 1;\mu)=\frac{\mu^{r_i}\exp(-\mu)}{r_i!\left\{1-\exp(-\mu)\right\}}\ ,r_i=1,\dots\ \ .$$

where $\mu=10$. The R_i inspection times $0<a_{i1}<\cdots<a_{iR_i}<1$ are then uniformly distributed over $[0,C]$. The number of events occuring between assessments are then $m_{ir}=\sum_{j=1}^{n_i}I(a_{i,r-1}<t_{ij}\leq a_{i,r})$, $r=1,\dots,R_i$. We consider a sample size of $m=500$

		EM-MLE				EM-TS			
PIECE	PARAMETER	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
		PHASE I: ONSET OF DAMAGE							
$[0.00, 5.38)$	α_{11}	-0.014	0.447	0.431	94.6	-0.005	0.475	0.457	93.8
$[5.38, 13.05)$	α_{12}	0.008	0.445	0.452	95.6	0.004	0.484	0.496	95.4
$[13.05, 25.26)$	α_{13}	0.021	0.413	0.427	95.6	0.008	0.447	0.458	96.4
$[25.26, 50.00)$	α_{14}	0.042	0.450	0.440	94.4	0.052	0.456	0.451	94.4
	β_{11}	0.968	6.491	6.494	95.4	1.000	6.449	6.538	95.6
	β_{12}	0.654	6.431	6.447	96.6	0.621	6.467	6.488	97.2
		PHASE II: PROGRESSION OF DAMAGE							
$[0.00, 9.23)$	α_{21}	-0.163	1.626	1.622	95.2	-0.167	1.626	1.623	95.0
$[9.23, 18.46)$	α_{22}	0.031	1.763	1.815	95.6	0.030	1.760	1.815	95.6
$[18.46, 27.68)$	α_{23}	0.104	2.071	1.994	94.8	0.103	2.068	1.994	94.8
$[27.68, 50.00)$	α_{24}	-0.003	1.781	1.702	93.2	-0.003	1.780	1.702	93.2
	β_{21}	-0.117	1.669	1.691	95.4	-0.118	1.669	1.691	95.2
	β_{22}	0.039	1.617	1.686	96.4	0.037	1.617	1.686	96.4

† EBIAS, ESE and ASE reported are $\times 10^2$

Table 4.1: Empirical performance† of estimators; sample size $m = 500$, number of simulations $nsim = 500$, $\alpha_1 = 0.036$, $\alpha_2 = 0.5$, $\beta_1 = (0.5, 0.5)$, $\beta_2 = (-0.5, -0.5)$; ASE are average of standard errors estimated via methods in Section 4.3.3 (EM-MLE) and Section 4.4.2 (EM-TS).

and simulate five hundred datasets ($nsim = 500$). For each dataset, we fit the proposed two-phase model by the EM algorithm under both simultaneous (maximum likelihood) and two-stage estimation; these are denoted by EM-MLE and EM-TS in Table 1 respectively. The break-points for both phases are chosen to correspond to the quartiles of the baseline survival function. Standard errors for the maximum likelihood estimators were obtained by Louis (Louis, 1982) (see Section 4.3.3) and using estimating function theory (see Section 4.4.2). The empirical coverage probabilities were computed as the proportion of all simulated datasets for which the 95% confidence interval contained the true parameter value.

The empirical performance of the estimators using both estimation approaches are shown in Table 4.1, where the empirical biases (EBIAS) are generally small. There is good agreement between the empirical standard errors (ESE) and average standard errors (ASE) obtained by Louis (Louis, 1982) or the methods of Section 4.4.2 respectively and the empirical coverage probabilities (ECP) are all compatible with the nominal level. From the simulation results, we can conclude that both estimation approaches give good performance; the empirical biases are relatively small and the empirical coverage probabilities are all compatible with the nominal 95% level. There is relatively little price to pay in terms of efficiency when the two-stage estimation procedure is used over maximum likelihood estimation.

4.5.2 Application of Psoriatic Arthritis Data

Here we consider the data on joint damage in patients with psoriatic arthritis from the University of Toronto Psoriatic Arthritis Registry. Specific interest lies in examining the effects of human leukocyte antigen (HLA) markers on the duration of the indolent phase

following diagnosis and on the rate of joint damage following the end of the indolent phase. The break-points for the model of the duration of the indolent phase are 3.5, 9.2, 13.7 and 26 years, derived from the nonparametric estimate of the cumulative probability function for the time to the precipitating event (first joint known to become damaged). The break-points for the second part were taken as 8.2, 12.6, 17.0 and 23.5 years likewise derived from the mean function for the cumulative number of events estimated by isotonic regression. We examine the effects of HLA markers selected based on the results of Wu and Cook (Wu and Cook, 2015), while controlling for gender and patient age.

As in the empirical studies, we find from the results in Table 4.2 that there is good agreement in the estimates obtained by the EM-MLE and EM-TS algorithms. We therefore discuss the results of maximum likelihood estimation here. Among the HLA markers, HLA-A11, HLA-A25, HLA-A29, HLA-A30, HLA-C03 and HA-DRB1-10 had insignificant association with the duration of the indolent phase but their presence was associated with a significant reduction of the rate of damage in the active phase of the disease.. For HLA-A11 for example, the relative rate of damage in the active phase associated with the presence of HLA-A11 is $RR = 0.70$ (95% CI : 0.53, 0.91; $p = 0.0087$); the corresponding relative rates for the other markers were HLA-A25 $RR = 0.07$ (95% CI : 0.01, 0.53; $p = 0.0096$), HLA-A29 $RR = 0.25$ (95% CI : 0.15, 0.43; $p < 0.0001$), HLA-A30 $RR = 0.78$ (95% CI : 0.63, 0.97; $p = 0.0235$), HLA-C03 $RR = 0.57$ (95% CI : 0.47, 0.70; $p < 0.0001$), and HLA-DRB1-10 $RR = 0.04$ (95% CI : 0.01, 0.27; $p = 0.0011$). Moreover, there is significant evidence that the effect of HLA-A25, HLA-A30, HLA-C03, and HLA-DRB1-10 on the duration of the indolent phase and on damage progression are different; see the last column of Table 2 for the homogeneity $p-$values. HLA-B27 is a known risk factor for disease progression in PsA and here we find its presence is associated with both a shorter indolent phase and more rapid disease progression; the same can be said for HLA-DQB1-02.

		PHASE I			*PHASE II*			
HLA Marker	METHOD	EST	SE	p	EST	SE	p	$p^{\dagger}$
HLA-A11	EM-MLE	-0.280	0.274	0.3079	-0.363	0.138	0.0087	0.7876
	EM-TS	-0.315	0.274	0.2506	-0.358	0.139	0.0098	0.8905
HLA-A25	EM-MLE	-0.211	0.597	0.7242	-2.627	1.014	0.0096	0.0407
	EM-TS	-0.159	0.597	0.7900	-2.614	1.015	0.0100	0.0378
HLA-A29	EM-MLE	-0.597	0.371	0.1076	-1.377	0.270	< 0.0001	0.0899
	EM-TS	-0.601	0.371	0.1055	-1.375	0.270	< 0.0001	0.0925
HLA-A30	EM-MLE	0.458	0.295	0.1208	-0.250	0.110	0.0235	0.0256
	EM-TS	0.364	0.299	0.2243	-0.249	0.110	0.0244	0.0564
HLA-B27	EM-MLE	0.468	0.183	0.0105	0.235	0.067	0.0004	0.2333
	EM-TS	0.490	0.183	0.0074	0.237	0.067	0.0004	0.1957
HLA-C03	EM-MLE	0.014	0.219	0.9480	-0.563	0.103	< 0.0001	0.0178
	EM-TS	0.017	0.220	0.9367	-0.566	0.103	< 0.0001	0.0169
HLA-C04	EM-MLE	-0.012	0.224	0.9576	-0.120	0.106	0.2565	0.6650
	EM-TS	0.011	0.225	0.9606	-0.122	0.107	0.2568	0.5985
HLA-DQB1-02	EM-MLE	0.386	0.164	0.0187	0.249	0.062	< 0.0001	0.4365
	EM-TS	0.394	0.164	0.0163	0.252	0.063	< 0.0001	0.4215
HLA-DRB1-10	EM-MLE	0.203	0.594	0.7326	-3.280	1.002	0.0011	0.0028
	EM-TS	0.195	0.594	0.7428	-3.275	1.002	0.0011	0.0029

$p^{\dagger}$ a p-value from a test of homogeneity.

Table 4.2: Results of fitting piecewise constant baseline hazard model for the duration of the indolent period and piecewise constant baseline rate model for the occurrence of joint damages under simultaneous (EM-MLE) and two-stage (EM-TW) estimation; $p-$values are based on Wald tests.

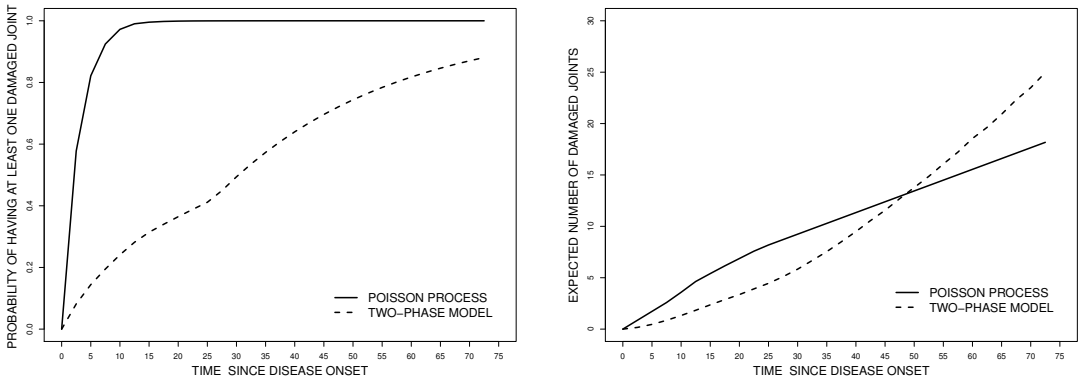


Figure 4.3: Estimates of the probability of having at least one damaged joint as measured from the time since disease onset (left panel) and the expected number of damaged joints for the time since disease onset (right panel).

4.6 Discussion

There are several avenues for generalizations of this work including the use of semiparametric models for the time from disease onset to the event signalling the beginning of the second phase of the process. Much work has been done in the last twenty years on the development of flexible regression methodology and statistical theory for the analysis of interval-censored failure time data (Sun, 2006). A more challenging generalization would be to relax the Markov assumption for the second phase process given the intermittent observation process. In the general multistate formulation of Section 4.2 much work has been done on methods for fitting and assessing Markov models in this setting but semi-Markov and models with hybrid time scales have seen little development. Mixed effect models which are Markov conditional on latent random effects however, have been developed and render more elaborate dependencies on the process history. The conditional Markov property enables one to fit these models even when the historical information is unobserved

due to the intermittent observation process. For the recurrent event model we consider in Section 4.3 this would correspond to a mixed Poisson model for the second phase of the process which would be negative binomial if a gamma distributed random effect were introduced. Large data sets would be required to estimate parameters in this more flexible model.

We have carried out tests of the null hypothesis of common coefficients for the phase I and II regression models. It would also be of interest to assess whether there is evidence of a need for the two-phase model because estimation, inferences and model interpretation would be so much easier if the second phase model were adequate. Such a test would be analogous to tests for the need to accommodate a non-susceptible fraction in cure rate models, but in this context this is more challenging since the timescale for the second phase model is defined as the time from the precipitating event.

Identification of important genetic and soluble biomarkers is of primary interest in psoriatic arthritis and the two-phase model offers an important opportunity to identify factors that may be prognostic for different aspects of the disease process. Given that a marker may be entertained in both phases one could consider the use of the group LASSO (Yuan and Lin, 2005; Wang and Leng, 2008) by defining pairs of coefficients for each marker, with one coefficient defined in the regression model for the phase I duration and another defined in the phase II model.

Often cohort data are created from registries which required individuals to have experienced some disease manifestation for enrolment. This can lead to a biased sampling scheme arising due to truncation of the disease process. Researchers may require individuals to not have experienced disease activity or damage to be eligible for an inception cohort, which would result in right-truncated interval-censored duration times for the first phase. Cohorts of individuals with advanced disease may require progression to some ad-

vanced state of the second phase process yielding right-truncated phase I and II data. The expectation-maximization algorithm we describe can be adapted to accommodate left-, right- and interval-truncation by the conceptualization of "ghosts" in the spirit of Turnbull (Turnbull, 1976). Such a complete data likelihood will be possible to fit with penalty terms using standard software for penalized Poisson regression.

Appendix 4.A Evaluation of $Q(\theta;\theta^{(v)})$ for Maximum Likelihood Estimation

4.A.1 Details of EM Algorithm

Here we show the details of the EM algorithm we described in Section 3.2. At vth iteration, we proceed as follows:

1. Evaluate $\tilde{\iota}_k^{(v)} = E[I_k(t_1)|D;\theta^{(v)}]$ and $\widetilde{\omega}_k^{(v)} = E[W_k(t_1)|D;\theta^{(v)}]$. Then

$$Q_1(\theta_1;\theta^{(v)}) = \delta_1 \left\{ \sum_{k=1}^{K_1} \tilde{\iota}_k^{(v)} \left(\log \alpha_{1k} + x'\beta_1\right) - \sum_{k=1}^{K_1} \alpha_{1k} \widetilde{\omega}_k^{(v)} e^{x'\beta_1} \right\} \\ -(1-\delta_1) \left(\sum_{k=1}^{K_1} \alpha_{1k} W_k(a_{R_1}) e^{x'\beta_1} \right) . \tag{4.A.1}$$

2. Maximize $Q_1(\theta_1;\theta^{(v)})$ to get the updated estimate of θ_1, $\theta_1^{(v+1)}$.
   Let $Z_k = (Z_{k1},\ldots,Z_{kK_1})'$ denote the indicator function, where $Z_{k\ell} = I(k=\ell), \ell = 1,\ldots,K_1$. Let $\alpha_k = \log \alpha_{1k}$, $k = 1,\ldots,K_1$ and $\alpha = (\alpha_1,\ldots,\alpha_{K_1})'$, then we can write

$$Q_1(\theta;\theta^{(v)}) = \sum_{k=1}^{K_1} \left[\delta_1 \left\{ \tilde{\iota}_k^{(v)} \left(z_k'\alpha + x'\beta_1\right) - \widetilde{\omega}_k^{(v)} e^{z_k'\alpha + x'\beta_1} \right\} - (1-\delta_1) W_k(a_{R_1}) e^{z_k'\alpha + x'\beta_1} \right] . \tag{4.A.2}$$

   We note that (4.A.2) has a Poisson form of log likelihood function, then we can use existing software (`glm`) to maximize it by creating a pseudo-dataset in the following format, as shown in Table 4.A.1.

3. Evaluate $\widetilde{u}_{rk}^{(v)} = E[u_{rk}|D;\theta_1^{(v+1)},\theta_2^{(v)}]$ and $\widetilde{n}_{rk}^{(v)} = n_r E[\alpha_{2k}^{(v)} u_{rk} / \sum_{k=1}^{K_2} \alpha_{2k}^{(v)} u_{rk}|D;\theta_1^{(v+1)},\theta_2^{(v)}]$.

ID (i)	piece (k)	Z_{k1}	Z_{k2}	$\cdots$	Z_{kK_1}	X_1	$\cdots$	X_p	Response	Offset
					$\delta_1 = 1$					
i	1	1	0	$\cdots$	0	x_1	$\cdots$	x_p	$\widetilde{\iota}_1^{(v)}$	$\log \widetilde{\omega}_1^{(v)}$
i	2	0	1	$\cdots$	0	x_1	$\cdots$	x_p	$\widetilde{\iota}_2^{(v)}$	$\log \widetilde{\omega}_2^{(v)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
i	K_1	0	0	$\cdots$	1	x_1	$\cdots$	x_p	$\widetilde{\iota}_{K_1}^{(v)}$	$\log \widetilde{\omega}_{K_1}^{(v)}$
					$\delta_1 = 0$					
i	1	1	0	$\cdots$	0	x_1	$\cdots$	x_p	0	$\log W_1(a_{R_1})$
i	2	0	1	$\cdots$	0	x_1	$\cdots$	x_p	0	$\log W_2(a_{R_1})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
i	K_1	0	0	$\cdots$	1	x_1	$\cdots$	x_p	0	$\log W_{K_1}(a_{R_1})$

Table 4.A.1: Pseudo-dataframe for the maximization of Q_1.

Then

$$
\begin{aligned}
& Q_2(\theta_2; \theta_1^{(v+1)}, \theta_2^{(v)}) \\
=& E_{T_1} E_{N_{rk}|T_1} \left[\log L_{\mathrm{C2}}(\theta_2) | D; \theta_1^{(v+1)}, \theta_2^{(v)} \right] \\
=& \delta_1 E \left[\sum_{r=1}^{R_2} \sum_{k=1}^{K_2} \left\{ n_r \frac{\alpha_{2k}^{(v)} u_{rk}}{\sum_{k=1}^{K_2} \alpha_{2k}^{(v)} u_{rk}} \left(\log \alpha_{2k} + x' \beta_2 \right) - \alpha_{2k} u_{rk} \exp(x' \beta_2) \right\} | D; \theta_1^{(v+1)}, \theta_2^{(v)} \right] \\
=& \sum_{r=1}^{R_2} \sum_{k=1}^{K_2} \delta_1 \left\{ \widetilde{n}_{rk}^{(v)} \left(\log \alpha_{2k} + x' \beta_2 \right) - \alpha_{2k} \widetilde{u}_{rk}^{(v)} \exp(x' \beta_2) \right\} .
\end{aligned}
\tag{4.A.3}
$$

4. Maximize $Q_2(\theta; \theta_1^{(v+1)}, \theta_2^{(v)})$ to get the updated estimate of θ_2, $\theta_2^{(v+1)}$. Let $Z_k = (Z_{k1}, \ldots, Z_{kK_2})'$ denote the indicator function, where $Z_{k\ell} = I(k = \ell), \ell = 1, \ldots, K_2$. Let $\gamma_k = \log \alpha_{2k}$, $k = 1, \ldots, K$ and $\gamma = (\gamma_1, \ldots, \gamma_{K_1})'$, then we can write

$$Q_2(\theta; \theta_1^{(v+1)}, \theta_2^{(v)}) = \sum_{r=1}^{R_2} \sum_{k=1}^{K_2} \delta_1 \left\{ \widetilde{n}_{rk}^{(v)} \left(z_k' \gamma + x' \beta_2 \right) - \widetilde{u}_{rk}^{(v)} \exp(z_k' \gamma + x' \beta_2) \right\} . \quad (4.A.4)$$

We note that (4.A.4) has a Poisson form of log likelihood function, then we can use existing software (`glm`) to maximize it by creating a pseudo dataset in the following format, as shown in Table 4.A.2.

ID (i)	assess (r)	piece (k)	Z_{k1}	Z_{k2}	$\cdots$	Z_{kK_2}	X_1	$\cdots$	X_p	Response	Offset
i	1	1	1	0	$\cdots$	0	x_1	$\cdots$	x_p	$\widetilde{n}_{11}^{(v)}$	$\log \widetilde{u}_{11}^{(v)}$
i	1	2	0	1	$\cdots$	0	x_1	$\cdots$	x_p	$\widetilde{n}_{12}^{(v)}$	$\log \widetilde{u}_{12}^{(v)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	
i	1	K_2	0	0	$\cdots$	1	x_1	$\cdots$	x_p	$\widetilde{n}_{1K_2}^{(v)}$	$\log \widetilde{u}_{1K_2}^{(v)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$				
i	R_2	1	1	0	$\cdots$	0	x_1	$\cdots$	x_p	$\widetilde{n}_{R_2,1}^{(v)}$	$\log \widetilde{u}_{R_2,1}^{(v)}$
i	R_2	2	0	1	$\cdots$	0	x_1	$\cdots$	x_p	$\widetilde{n}_{R_2,2}^{(v)}$	$\log \widetilde{u}_{R_2,2}^{(v)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	
i	R_2	K_2	0	0	$\cdots$	1	x_1	$\cdots$	x_p	$\widetilde{n}_{R_2,K_2}^{(v)}$	$\log \widetilde{u}_{R_2,K_2}^{(v)}$

Table 4.A.2: Pseudo-dataframe for the maximization of Q_2.

4.A.2 Evaluations of the Conditional Expectations in the E-Step

Since all the unobserved quantities are related to T_1, we need conditional expectations in the form of

$$\int_{L1}^{R_1} f(x)\mathrm{d}x \ .$$

Due to the complicated nature of the function $f(x)$, closed form expressions are not available so we use numerical integration. Here we describe the Gaussian Quadrature which we used in the analyses.

First we can use a linear transformation to change this integration into a new integration on the interval $(-1, 1)$.
Let $x = \phi(y) = \{y(R1 - L1) + L_1 + R_1\}/2$, so

$$\int_{L1}^{R_1} f(x)\mathrm{d}x = \int_{-1}^{1} f(\phi(y))\phi'(y)\mathrm{d}y \ .$$

Then using Chebyshev quadrature (Golub and Welsch, 1969) of the 1st kind with the weight function $w(y) = 1/\sqrt{1-y^2}$, we approximate the integration by:

$$\begin{aligned}\int_{L1}^{R_1} f(x)\mathrm{d}x = \int_{-1}^{1} f(\phi(y))\phi'(y)\mathrm{d}y &= \int_{-1}^{1} w(y)\frac{f(\phi(y))\phi'(y)}{w(y)}\mathrm{d}y \\ &\doteq \int_{-1}^{1} w(y)g(y)\mathrm{d}y = \sum_{s=1}^{N} w_s g(y_s) \ ,\end{aligned}$$

where w_s and y_s are the weights and nodes that are picked based on weight function $w(y)$. Monte Carlo methods with rejection sampling could alternatively be used to approximate these expectations by simulation.

Chapter 5

Discussion and Future Research

5.1 Penalized Regression for Interval-Censored Times

Much of the work on variable selection techniques was initially carried out in the context of continuous responses, but advances have been made to deal with binary responses and time to event responses. For the latter, when times are right-censored, the penalty term is typically applied to the partial likelihood arising from a semiparametric Cox regression model. Many prospective studies, however, involve event times subject to interval censoring (Sun, 2006). In cancer clinical trials, for example, new metastatic lesions are often only detectable by imaging (Hortobagyi et al., 1996), so the time from randomization to the development of a new lesion is unknown. As another example, in patients infected with cytomegalovirus, the time from infection to viral shedding in the blood is only known to lie between the last negative and first positive serum sample (Betensky and Finkelstein, 1999). Finally, the occurrence of an asymptomatic fracture in osteoporosis patients is only detected by radiographic examination (Riggs et al., 1990).

In Chapter 2, we considered the problem of variable selection in the context of interval-censored time to event data. We adopt a flexible piecewise exponential model (Friedman, 1982) for the event of interest and penalize the complete data likelihood constructed by treating the interval-censored failure times as known. An expectation-maximization (EM) algorithm (Dempster et al., 1977) is then used for variable selection through optimization of the observed data likelihood incorporating the LASSO, adaptive LASSO or SCAD penalty function.

Important topics of future work include use of more flexible semiparametric methods in this setting, including methods based on local likelihood (Betensky et al., 2002) or penalized splines (Cai and Betensky, 2003). The properties of coefficients obtained following variable selection are not well understood, but Lockhart et al. (2014) represents a recent advance. Derivation of the limiting behaviours of estimators resulting from semiparametric models will be important.

A natural extension of this work is for the analysis of recurrent events observed subject to interval-censoring. In the psoriatic arthritis clinic, when interest lies in modeling the cumulative number of damaged joints, this count is often based upon damage scores determined only upon radiographic examination. The resulting data, consisting of a series of assessment times and counts representing the number of events occurring between consecutive assessments, is often called panel count data (Sun and Kalbfleisch, 1995). Lawless and Zhan (1998) develop the likelihood and estimating functions for the analysis of such data for mixed Poisson models with piecewise-constant rate functions (Cook and Lawless, 2007). The former can be naturally adapted to allow variable selection based on penalized likelihood for recurrent event data. Given the individual patient level random effect, the penalized likelihood has a similar form to the one we have in this setting. While the observed data likelihood can be penalized, a complete data likelihood involving a more

detailed recording of the counts and the patient level random effect is very appealing and could still exploit existing software. See He et al. (2009) for a semiparametric implementation of a similar algorithm. Tong et al. (2009) develop penalized estimating functions for variable selection with panel count data and (Wu and He, 2012) propose and study a fast and efficient coordinate ascent algorithm for the same problem.

5.2 Penalized Regression for Truncated and Censored Times

The availability of large disease registries with longitudinal follow-up has lead to increased interest in utilizing such data for scientific inquiry about the genetic basis for disease onset and progression. In disease processes with multiple stages it is often the case that one registry may require individuals to be in an early phase of a disease process while another may have selection criteria requiring individuals be in a more advanced stage.

There are many clinical trials and cohort studies which involve event times subject to truncation due to biased sampling schemes. There are three truncation schemes: left-truncation, right-truncation and interval-truncation. For example, the time from the onset of psoriasis (Ps) to the onset of psoriatic arthritis (PsA) is the event time of interest and two cohort studies are available to be used: psoriasis (Ps) and psoriatic arthritis (PsA) cohorts. In the Ps cohort, only patients who have been diagnosed of Ps before the study starts are included in the sample, then when the date of psoriasis onset is to be used as the time origin, this sample is left-truncated. In the PsA cohort, only patients who have been diagnosed of PsA before the study starts are included in the sample, so this yields right-truncation. In Section 2.6, we consider the problem of variable selection in the context

of truncated samples and propose an expectation-maximization algorithm to deal with truncated event times by using a Turnbull-type (Turnbull, 1976) complete data likelihood which involves the pseudo-individuals in the population who did not satisfy the truncation conditions and incorporating the LASSO, adaptive LASSO or SCAD penalty function. We describe a penalized EM algorithm based on a piecewise exponential response model, for which existing techniques for variable selection can be exploited to handle truncated event times.

Future work includes a natural extension of this work to deal with multistate process. For example, we consider another motivating problem involving the disease course in psoriatic arthritis, represented in the following multistate diagram. The states are numbered

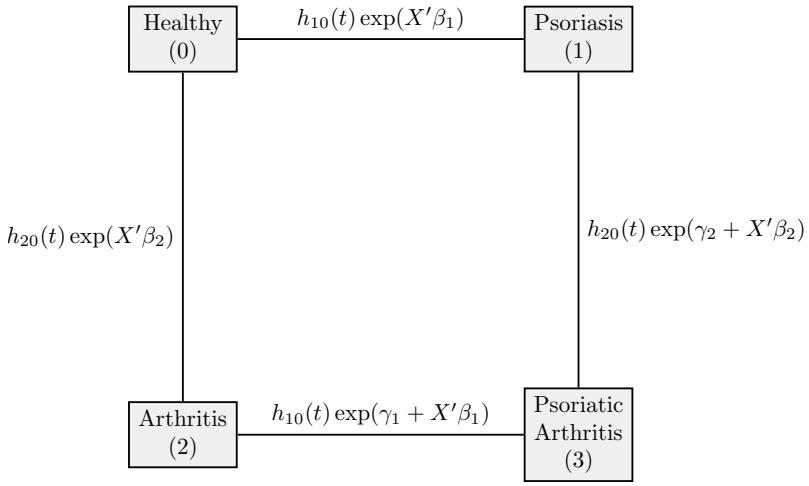


Figure 5.1: Multistate diagram illustrating the course of disease in psoriatic arthritis cohort.

from 0 to 3, with 0 denoting a healthy condition, 1 denoting psoriasis, 2 denoting arthritis and 3 denoting psoriatic arthritis. Patients are included into the PsA registry if they have developed psoriatic arthritis by the time they are screened; that is, they are in state 3 at time $\mathcal{R}$. While our focus in Chapter 2 was on the waiting time from psoriasis to psoriatic

arthritis necessarily focussing on those known to have developed psoriasis first, patients can develop either psoriasis or arthritis first. If one has psoriasis before arthritis then this patient makes transitions in the path $0 \to 1 \to 3$; if one has arthritis before psoriasis then this patient make transitions $0 \to 2 \to 3$ path.

Let T_1 denote the time of psoriasis and T_2 denote the time of arthritis, so the time to PsA is $\max(T_1, T_2)$. The observed data likelihood in this case is

$$L_{\text{obs}} \propto f_{T_1,T_2}(t_1, t_2 | \max(T_1, T_2) < \mathcal{R}, X) \ .$$

A Turnbull-type complete data likelihood is then of form

$$L_{\text{C}} \propto f_{T_1,T_2}(t_1, t_2|X) \times \prod_{r=0}^{2} P(Z(\mathcal{R}) = r|Z(0) = 0, z)^{J_{0r}} \ ,$$

where $Z(s)$ reflects the state occupied at time $s > 0$, J_{0r} is an unknown variable, which represents the number of "ghosts" who are in state r at time $\mathcal{R}$, $r = 0, 1, 2$. A penalized complete data log-likelihood can be defined to be

$$\frac{1}{m} \log L_{\text{C}}(\theta) - p_{\gamma,\lambda}(\beta) \ ,$$

where $p_{\gamma,\lambda}(\beta)$ is the penalty function with form

$$\lambda \sum_{j=1}^{p} \sqrt{\beta_{1j}^2 + \beta_{2j}^2} \ ,$$

$$(1-\gamma)\lambda \sum_{j=1}^{p} \sqrt{\beta_{1j}^2 + \beta_{2j}^2} + \gamma\lambda \sum_{j=1}^{p} \{|\beta_{1j}| + |\beta_{2j}|\}$$

for group LASSO and sparse group LASSO respectively. The estimation procedure can be done through creating a pseudo-dataset at each expectation step and maximizing the penalized likelihood by existing software of group LASSO and sparse group LASSO at each maximization step.

5.3 Assessing the Accuracy of Predictive Models with Interval-Censored Data

A related research project following the variable selection from Chapter 2 involved the development of techniques for assessing the predictive accuracy of models when interest lies in the time of an event which is interval-censored through use of inverse weighting probability of censoring techniques.

In the survival context, we obtain flexible prediction models and often evaluate their predictive value on the same set of data; or better still, based on an external validation data set. The purpose of assessing the predictive accuracy of a regression model is often to establish whether a prognostic model can be used to reliably predict patients event status at a particular time and to provide a basis for clinical decision making. There has been a lot of research conducted focusing on prediction with time to event data subject to right-censoring. Chapter 3 considers the challenge of assessing the accuracy of a predictive model when response times are interval-censored. Inverse probability weighted (IPW) and augmented IPW estimators are developed and evaluated based on the mean prediction error and the area under the receiver operating characteristic curve. The weights are estimated from multistate model which jointly considers the event, the inspection and the censoring processes.

We remark that with an independent validation sample one could retain the prediction model obtained from a training dataset but based weights on a new inspection model, censoring model and a new response model. In this case if the response model obtained based on the validation sample is correct and leads to correct specification of the weights, then the double robustness property can be realized. That is, if the prediction model from the training sample is correct a consistent estimator of the prediction error is obtained, and

alternatively if the model for the weights selected independently based on the validation data is correct, a consistent estimator of the prediction error is obtained. As an independent validation sample is not available at this point we do not explore this double robustness property further here but it represents an interesting and important area of future research.

In future research, we also plan to consider the case of truncated and interval-censored data. While we have dealt with the interval-censored data here, it is less clear how one might assess predictive accuracy when samples are chosen subject to truncation, but this feature is often present in problems involving large datasets. Another avenue of research to generalize this work is to consider taking the inverse probability weighting into consideration along with variable selection to adjust for the effect of informative inspection process.

5.4 Statistical Models for Complex Life History Processes

Many chronic disease processes feature considerable variability in their course which must be dealt with in statistical analysis for valid inference. Regression modeling and regression diagnostics play a central role in explaining this variation in such a way that scientific understanding can be advanced. There are lots of work that have been done to accommodate such variations; for example, mixture models and cure rate models. In Chapter 4, we formulated a two-phase model in which phase I incorporates an initial time from disease onset to the commencement of phase II, and a second part which characterizes the nature of the process during phase II. This model can be used to separately examine prognostic factors for the length of the inactive phase as well as factors prognostic for the

nature and rate of change, for example, in the active phase. In some settings this will offer a more appropriate representation of complex multi-phase disease processes, can help identify different types of risk factors, and could yield more accurate prediction models.

Use of semiparametric models for the time from disease onset to the event signalling, the beginning of the second phase of the process is one way to generalize this work. There has been much work on the development of flexible regression methodology and statistical theory for the analysis of interval-censored failure time data (Sun, 2006). A more challenging generalization would be to relax the Markov assumption for the second phase process given the intermittent observation process.

One primary interest in the psoriatic arthritis cohort study is to identify important genetic and soluble biomarkers associated with the disease progression. The proposed two-phase model offers an important opportunity to identify factors that may be prognostic for different aspects of the disease process. Given that a marker may be entertained in both parts of the model one could consider the use of the group LASSO (Yuan and Lin, 2005; Wang and Leng, 2008) by defining pairs of coefficients for each marker, with one coefficient defined in the regression model for the phase I duration and another defined in the phase II model. Variations of this such as the sparse group LASSO (Simon et al., 2013) could be useful in selecting variables by group and within group.

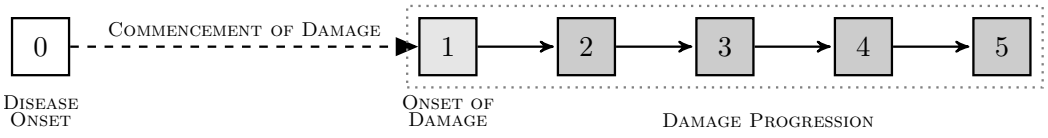


Figure 5.2: Multistate diagram illustrating the two phases of the disease process.

A penalized complete data log-likelihood can be defined to be

$$\frac{1}{m} \log L_{\mathrm{C}}(\theta) - p_{\gamma,\lambda}(\beta) \ ,$$

where $\log L_C(\theta)$ is in the form of (4.10) and $p_{\gamma,\lambda}(\beta)$ is the penalty function for group LASSO and sparse group LASSO. The estimation procedure can be done through using the EM algorithm proposed in Chapter 4 along with using existing software of group LASSO and sparse group LASSO at each maximization step.

Researchers may require individuals to not have experienced disease activity or damage to be eligible for an inception cohort, which would result in right-truncated interval-censored duration times for the first phase. Cohorts of individuals with advanced disease may require progression to some advanced state of the second phase process yielding right-truncated phase I and II data. The expectation-maximization algorithm we describe can be adapted to accommodate left-, right- and interval-truncation by the conceptualization of "ghosts" in the spirit of Turnbull (Turnbull, 1976). Such a complete data likelihood will be possible to fit with penalty terms using standard software for penalized Poisson regression.

Finally, the more general two-phase model accommodates a more complex disease process and as a consequence one might expect it to perform better in terms of prediction of outcomes. Use of the model- or imputation-based procedures for estimating the prediction error would be good to investigate in simulation, along with the inverse probability weighted and augmented inverse probability weighted estimators of prediction error.

References

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, 8(8):907–925.

Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics*, pages 1299–1327.

Altman, D. G., McShane, L. M., Sauerbrei, W., and Taube, S. E. (2012). Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. *BMC Medicine*, 10(1):51.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (2012). *Statistical Models Based on Counting Processes*. Springer Science & Business Media, New York.

Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11(2):91–115.

Betensky, R. A. and Finkelstein, D. M. (1999). A non-parametric maximum likelihood estimator for bivariate interval censored data. *Statistics in Medicine*, 18(22):3089–3100.

Betensky, R. A., Lindsey, J. C., Ryan, L. M., and Wand, M. P. (2002). A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine*, 21:263–275.

Bradic, J., Fan, J., and Jiang, J. (2011). Regularization for Coxs proportional hazards model with NP-dimensionality. *Annals of Statistics*, 39(6):3092–3120.

Braun, J., Duchesne, T., and Stafford, J. E. (2005). Local likelihood density estimation for interval censored data. *The Canadian Journal of Statistics*, 33:39–60.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24(6):2350–2383.

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized splines. *Biometrics*, 59(3):570–579.

Chandran, V., Cook, R. J., Edwin, J., Shen, H., Pellett, F. J., Shanmugarajah, S., Rosen, C. F., and Gladman, D. D. (2010). Soluble biomarkers differentiate patients with psoriatic arthritis from those with psoriasis without arthritis. *Rheumatology*, 49(7):1399–1405.

Chandran, V., Cook, R. J., Thavaneswaran, A., Lee, K.-A., Pellett, F., and Gladman, D. (2012). Parametric survival analysis as well as multi-state analysis confirms the association between human leukocyte antigen alleles and the development of arthritis mutilans in patients with psoriatic arthritis. *Journal of Rheumatology*, 39(8):1723–1723.

Cook, R. J. and Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events*. Springer, New York.

Cook, R. J. and Lawless, J. F. (2014). Statistical issues in modeling chronic disease in cohort studies. *Statistics in Biosciences*, 6(1):127–161.

Cook, R. J., Zeng, L., and Lee, K.-A. (2008). A multistate model for bivariate interval-censored failure time data. *Biometrics*, 64(4):1100–1109.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470.

Efron, B. (2004). The estimation of prediction error. *Journal of the American Statistical Association*, 99:619–632.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*, volume 57. CRC press.

Fan, J., Feng, Y., Samworth, R., and Wu, Y. (2010). *SIS: Sure Independence Screening*. R package version 0.6.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics*, 30(1):74–99.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, 14(3):257–262.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *Annals of Statistics*, 10:101–113.

Frydman, H. (1984). Maximum likelihood estimation in the mover-stayer model. *Journal of the American Statistical Association*, 79(387):632–638.

Gerds, T. A. and Schumacher, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040.

Gerds, T. A. and Schumacher, M. (2007). Efron-type measures of prediction error for survival analysis. *Biometrics*, 63(4):1283–1287.

Gladman, D. D., Schentag, C. T., Tom, B. D. M., Chandran, V., Brockbank, J., Rosen, C., and Farewell, V. T. (2008). Development and initial validation of a screening questionnaire for psoriatic arthritis: the Toronto Psoriatic Arthritis Screen (ToPAS). *Annals of the Rheumatic Diseases*, 68(4):497–501.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Golub, G. H. and Welsch, J. H. (1969). Calculation of gauss quadrature rules. *Mathematics of Computation*, 23(106):221–230.

Goodman, L. A. (1961). Statistical methods for the mover-stayer model. *Journal of the American Statistical Association*, 56(296):841–868.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545.

Grüger, J., Kay, R., and Schumacher, M. (1991). The validity of inferences based on incomplete observations in disease state models. *Biometrics*, 47:595–605.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

He, X., Tong, X., and Sun, J. (2009). Semiparametric analysis of panel count data with correlated observation and follow-up times. *Lifetime Data Analysis*, 15(2):177–196.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344.

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hogan, J. W., Roy, J., and Korkontzelou, C. (2004). Handling drop-out in longitudinal studies. *Statistics in Medicine*, 23(9):1455–1497.

Hortobagyi, G. N., Theriault, R. L., Porter, L., Blayney, D., Lipton, A., Sinoff, C., Wheeler, H., Simeone, J. F., Seaman, J., and Knight, R. D. (1996). Efficacy of pamidronate in reducing skeletal complications in patients with breast cancer and lytic bone metastases. *New England Journal of Medicine*, 335(24):1785–1792.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, 7(3):355–373.

Klein, J. P. (1992). Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, pages 795–806.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York.

Korn, E. L. and Simon, R. (1990). Measures of explained variation for survival data. *Statistics in Medicine*, 9(5):487–503.

Kyle, S., Chandler, D., Griffiths, C. E. M., Helliwell, P., Lewis, J., McInnes, I., Oliver, S., Symmons, D., and McHugh, N. (2005). Guideline for anti-TNF-α therapy in psoriatic arthritis. *Rheumatology*, 44(3):390–397.

Lawless, J. and Zhan, M. (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *Canadian Journal of Statistics*, 26(4):549–565.

Lawless, J. F. and Yuan, Y. (2010). Estimation of prediction error for survival models. *Statistics in Medicine*, 29(2):262–274.

Li, J. and Ma, S. (2013). *Survival Analysis in Medicine and Genetics*. CRC Press.

Lindsey, J. C. and Ryan, L. M. (1998). Tutorial in biostatistics: methods for interval-censored data. *Statistics in Medicine*, 17:219–238.

Lockhart, R., Taylor, J., Tibshirani, R. J., Tibshirani, R., et al. (2014). A significance test for the LASSO. *Annals of Statistics*, 42(2):413–468.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233.

Mallows, C. L. (1973). Some comments on c_p. *Technometrics*, 15(4):661–675.

Martinussen, T. and Scheike, T. H. (2007). *Dynamic Regression Models for Survival Data*. Springer Science & Business Media, New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall, 2nd edition.

McShane, L. M., Altman, D. G., and Sauerbrei, W. (2005a). Identification of clinically useful cancer prognostic factors: what are we missing? *Journal of the National Cancer Institute*, 97(14):1023–1025.

McShane, L. M., Altman, D. G., Sauerbrei, W., Taube, S. E., Gion, M., Clark, G. M., et al. (2005b). Reporting recommendations for tumor marker prognostic studies (REMARK). *Journal of the National Cancer Institute*, 97(16):1180–1184.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245.

Nielsen, G. G., Gill, R. D., Andersen, P. K., and Sørensen, T. I. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, pages 25–43.

Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677.

Polley, M.-Y. C., Freidlin, B., Korn, E. L., Conley, B. A., Abrams, J. S., and McShane, L. M. (2013). Statistical and practical considerations for clinical evaluation of predictive biomarkers. *Journal of the National Cancer Institute*, 105(22):1677–1683.

Preisser, J. S., Lohman, K. K., and Rathouz, P. J. (2002). Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, 21(20):3035–3054.

Queiro, R., Torre, J. C., González, S., López-Larrea, C., Tinturé, T., and López-Lagunas, I. (2003). HLA antigens may influence the age of onset of psoriasis and psoriatic arthritis. *The Journal of Rheumatology*, 30(3):505–507.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahman, P., Gladman, D. D., Cook, R. J., Zhou, Y., Young, G., and Salonen, D. (1998). Radiological assessment in psoriatic arthritis. *Rheumatology*, 37(7):760–765.

Rahman, P., Roslin, N. M., Pellett, F. J., Lemire, M., Greenwood, C. M., Beyene, J., Pope, A., Peddle, L., Paterson, A. D., Uddin, M., et al. (2011). High resolution mapping in the major histocompatibility complex region identifies multiple independent novel loci for psoriatic arthritis. *Annals of the Rheumatic Diseases*, 70(4):690–694.

Riggs, B. L., Hodgson, S. F., O'Fallon, W. M., Chao, E. Y., Wahner, H. W., Muhs, J. M., Cedel, S. L., and Melon, L. J. (1990). Effect of fluoride treatment on the fracture

rate in postmenopausal women with osteoporosis. *New England Journal of Medicine*, 322(12):802–809.

Schäfer, T. (2006). Epidemiology of psoriasis. *Dermatology*, 212(4):327–337.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer, New York.

Sun, J. and Kalbfleisch, J. D. (1995). Estimation of the mean function of point processes based on panel count data. *Statistica Sinica*, 5:279–290.

Therneau, T. M. (2013). A package for survival analysis in S. R package version 2.37-4.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. (1997). The LASSO method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused LASSO. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tom, B. D., Chandran, V., Farewell, V. T., Rosen, C. F., and Gladman, D. D. (2015). Validation of the Toronto Psoriatic Arthritis Screen Version 2 (ToPAS 2). *The Journal of Rheumatology*, 42(5):841–846.

Tong, X., He, X., Sun, L., and Sun, J. (2009). Variable selection for panel count data via non-concave penalized estimating function. *Scandinavian Journal of Statistics*, 36(4):620–635.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117.

Wang, H. and Leng, C. (2008). A note on adaptive group LASSO. *Computational Statistics and Data Analysis*, 52(12):5277–5286.

Winchester, R., Minevich, G., Steshenko, V., Kirby, B., Kane, D., Greenberg, D. A., and FitzGerald, O. (2012). HLA associations reveal genetic heterogeneity in psoriatic arthritis and in the psoriasis phenotype. *Arthritis & Rheumatism*, 64(4):1134–1144.

Witten, D. M. and Tibshirani, R. (2009). Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*, 19:29–51.

Wu, C. J. (1983). On the convergence properties of the em algorithm. *Annals of Statistics*, pages 95–103.

Wu, T. T. and He, X. (2012). Coordinate ascent for penalized semiparametric regression on high-dimensional panel count data. *Computational Statistics & Data Analysis*, 56(1):25–33.

Wu, Y. and Cook, R. J. (2015). Penalized regression for interval-censored times of disease progression: Selection of HLA markers in psoriatic arthritis. *Biometrics*, 71(3):782–791.

Wu, Y. and Cook, R. J. (2016). Variable selection and prediction in biased samples with censored outcomes. Submitted to Lifetime Data Analysis.

Yuan, M. and Lin, Y. (2005). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(1):49–67.

Yuan, Y. (2008). *Prediction Performance of Survival Models*. PhD thesis, University of Waterloo.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942.

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533.