

Variable Selection and Prediction for Multistate Processes under Complex Observation Schemes

by

Xianwei Li

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Statistics (Biostatistics)

Waterloo, Ontario, Canada, 2025

© Xianwei Li 2025

Examining Committee Membership

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

External Examiner:	Wenyu Jiang Professor Dept. of Mathematics and Statistics, Queen's University
Supervisor(s):	Richard J. Cook University Professor Dept. of Statistics and Actuarial Science, University of Waterloo Liqun Diao Associate Professor Dept. of Statistics and Actuarial Science, University of Waterloo
Internal Member:	Leilei Zeng Professor Dept. of Statistics and Actuarial Science, University of Waterloo Joel Dubin Professor Dept. of Statistics and Actuarial Science, University of Waterloo
Internal-External Member:	John Hirdes Professor School of Public Health Sciences, University of Waterloo

Author's Declaration

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

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

Statement of Contributions

A version of Chapter 2 of this thesis has been published in Statistics in Medicine. Chapter 3 has been prepared as a research paper and submitted to Statistics in Medicine, where it is currently under review. A manuscript based on Chapter 4 is in preparation for submission to a journal.

All three papers are co-authored with my thesis supervisors, and my contributions include formulating the research problems, deriving the main theorems, implementing the proposed methods in R, conducting simulations and real data analyses, and drafting the initial manuscripts.

Abstract

This thesis addresses variable selection and prediction in time-to-event analysis under complex observation schemes that commonly arise in biomedical studies. Such schemes may lead to right-censored data, interval-censored event times, or dual-censoring scenarios. Across three main chapters, we develop variable selection methods for multistate processes, address challenges arising from incomplete data under complex observation schemes, and investigate the implications of model misspecification, such as using simpler models in place of multistate models, and the potential risks of violating assumptions on covariate effects estimation and predictive performance.

We begin with considering the problem of variable selection for progressive multistate processes under intermittent observation in Chapter 2. This study is motivated by the need to identify which among a large list of candidate markers play a role in the progression of joint damage in psoriatic arthritis (PsA) patients. We adopted a penalized log-likelihood approach and developed an innovative Expectation-Maximization (EM) algorithm such that the maximization step can exploit existing software for penalized Poisson regression thereby enabling flexible use of common penalty functions. Simulation studies show good performance in identifying important markers with different penalty functions. We applied the algorithm in the motivating application involving a cohort of patients with psoriatic arthritis with repeated assessments of joint damage, and identified human leukocyte antigen (HLA) markers which are associated with disease progression, among a large group of candidate markers.

Chapter 3 extends this algorithm to more general multistate processes, and to more complex observation schemes. We consider the classical illness-death model which offers a useful framework for studying the progression of chronic disease while jointly modeling

death. The exact time of disease progression is not observed directly but progression status is recorded at intermittent assessment times; the time to death is subject to right-censoring. This creates a dual observation scheme where progression times are interval-censored and survival times are subject to right censoring. A penalized observed data likelihood approach is proposed which allows for separate penalties across different intensity functions. An EM algorithm is again developed to facilitate use of different penalties for variable selection on disease progression and death through penalized Poisson regression. This adaptation retains the flexibility to exploit existing software with commonly used penalty functions. Simulation studies show good finite-sample performance in variable selection with different combination of penalty functions. We also explored how various aspects of the variable selection algorithm affect performance such as use of nonparametric baseline intensities and different ways to select the optimal tuning parameter(s). An application to data from the National Alzheimer’s Coordinating Center (NACC) demonstrates the use of our method in jointly modeling dementia progression and mortality.

Chapter 4 builds on insights from Chapters 2 and 3 by investigating how simpler marginal methods targeting entry time to the absorbing state (e.g., a Cox proportional hazards model) compared to full multistate models. Here we retain use of the illness-death process as the basis of the investigation, but consider settings where transition times are only right-censored. We first study the limiting values of regression estimators from a Cox proportional hazards model when the data generating process is based on a Markov illness-death model. The potential impact of modeling the multistate processes based on a misspecified model is also investigated by considering cases where a) important covariates are omitted, or b) the Markov assumption is violated. We then examine the implications of model misspecification when the goal is prediction - this is done by evaluating the predictive performance of a misspecified Cox regression model for overall survival and

a misspecified Fine-Gray model for disease progression, and comparing their respective predictive performance against that of the true illness-death model. We find that the limiting value of regression coefficients estimators obtained from Cox models and Fine-Gray models depend on several factors, including the baseline hazard ratio of death between the intermediate and initial states, the probability of moving through the intermediate state, and covariate effects on all transitions. However, the corresponding predictive accuracy is not substantially compromised despite biases in the regression coefficient estimators in most scenarios we investigated. The limiting value of regression coefficients obtained from a Markov illness-death model and the corresponding predictive accuracy are sensitive to model misspecification such as omitting important covariates and violation of the Markov assumption. The practical implications are illustrated using a dataset of patients with metastatic breast cancer in the control arm to predict overall survival and fracture risk.

Chapter 5 reviews the contributions of this thesis and discusses problems warranting future research.

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to my supervisors, Dr. Richard Cook and Dr. Liqun Diao, for their patient guidance, encouragement, and support throughout my research journey. Under their mentorship, I grew from a novice into a confident researcher. Richard's passion for his work and care for his students were a constant source of my inspiration, and our meetings often felt like an energy station that recharged my motivation. Liqun's calm, clear, and logical way of presenting ideas and raising questions has greatly shaped my communication, presentation, and critical thinking. They listen attentively and are always willing to discuss not only research but also broader, interesting topics about life and perspectives. They have been mentors to me in far more than just research. I love spending time with them and am always excited to share more about myself, even the little things in my life.

I would also like to thank my committee members for their valuable feedback and support, and Mary Lou and Greg for their kindness and help in making our department such a warm and welcoming place. My sincere thanks go to Dr. Lihi Eder for the inspiring discussions and collaborative opportunities in psoriatic arthritis.

I am deeply grateful to Professor Ken Seng Tan, who interviewed and selected me for the exchange program in 2018. My academic journey in Canada would never have begun without him.

Special thanks go to Xiao and Alina, who helped me recognize my true self and realize my full potential. I am also grateful to my friends in Canada, especially Zijin, who has supported and accompanied me from the very beginning of this journey in a foreign land. I want to thank my office mate Alexandra and everyone in our lab - Laura, Aelita, Lily, Jiayi, Kecheng, Mu, Xiaoya, and Alexandra for creating such a warm and supportive community,

as well as the senior students Ce, Menglu, Jianchu, and Lijia for their generous and caring guidance.

Finally, I am thankful to Dingding, who can make anything funny and always brings laughter to my life; to my parents, for their unconditional support and their willingness to grow together with me; and to my grandmothers, who always make me feel like the best granddaughter in the world.

Dedication

To the memory of Xinhua Huang, who taught me integrity, responsibility, and kindness.

Table of Contents

Examining Committee	ii
Author’s Declaration	iii
Statement of Contributions	iv
Abstract	v
Acknowledgements	viii
Dedication	x
List of Figures	xvii
List of Tables	xxii
1 Introduction	1
1.1 Overview	1

1.2	Statistical Models and Methods for Lifetime Data	2
1.2.1	Multistate Processes	3
1.2.2	Multistate Processes under Complex Observation Scheme	5
1.2.3	Variable Selection for Time-to-event Analysis	6
1.3	Motivating Research Programs	8
1.3.1	The University of Toronto Psoriatic Arthritis Clinic (UTPAC)	8
1.3.2	The National Alzheimer’s Coordinating Center (NACC)	9
1.3.3	A Randomized Trial of Patients with Skeletal Metastases	9
1.4	Introduction to Research Topics	10
1.4.1	Variable Selection for Progressive Multistate Models	10
1.4.2	Variable Selection for Illness-death Processes under Dual Observa- tion Schemes	11
1.4.3	Model Misspecification in Multistate Analysis and Implications on Prediction	12
2	Variable Selection for Progressive Multistate Processes Under Intermit- tent Observation	13
2.1	Introduction	13
2.1.1	Background	13
2.1.2	HLA Markers and Sacroiliac Joint Damage in PsA	16
2.2	Variable Selection for Transition Intensities	18
2.2.1	Multistate Model and Likelihood Construction	18

2.2.2	The Penalized Observed Data Likelihood	21
2.3	Variable Selection via an Expectation-Maximization Algorithm	23
2.3.1	The Complete Data Log-likelihood	23
2.3.2	The Expectation-Maximization Algorithm	25
2.4	Simulation Studies	29
2.4.1	Design of Simulation Studies	30
2.4.2	Performance Metrics	33
2.4.3	Empirical Results	36
2.5	HLA Markers and Sacroiliac Joint Damage Data	40
2.6	Discussion	44
3	Variable Selection for Illness-death Processes under Dual Observation Schemes	47
3.1	Introduction	47
3.2	Motivating Problem and General Framework	51
3.2.1	Prognostic Factors of Dementia in National Alzheimer’s Coordinating Center (NACC) Cohort	51
3.2.2	The Penalized Observed Data Log-likelihood under a Dual Observation Scheme	54
3.3	A Penalized Expectation-Maximization Algorithm	58
3.3.1	Likelihood Construction for an Illness-Death Process under Continuous Observation	58

3.3.2	The Penalized Expectation-Maximization Algorithm	59
3.3.3	A Markov Formulation with Piecewise-Constant Baseline Intensities	60
3.3.4	Specification and Incorporation of Penalty Functions	64
3.4	Simulation Studies	67
3.4.1	Simulation Design	67
3.4.2	Evaluation Metrics	70
3.4.3	Empirical Results	72
3.5	Selecting Predictors for Dementia Onset and Death	78
3.6	Discussion	82
4	Impact of Model Misspecification with Multistate Processes: Some Results for Regression Analysis and Prediction	84
4.1	Introduction	84
4.1.1	Multistate and Alternative Models	84
4.2	Notation	87
4.2.1	Notation for General Multistate Processes	87
4.2.2	An Illness-death Process and Two Models	91
4.3	Limiting Values for Estimators of Regression Coefficients	93
4.3.1	Limiting Values of Cox Regression Estimators	93
4.3.2	Misspecified Multistate and Cox Models	98
4.4	Prediction Performance under Misspecified Multistate Models	103

4.4.1	Prediction of Overall Survival	105
4.4.2	Omission of an Important Covariate	113
4.4.3	Violation of the Markov Assumption	114
4.4.4	Prediction for Disease-Progression	114
4.5	Simulation Studies	122
4.6	Prediction of Fractures for Patients with Metastatic Breast Cancer	127
4.7	Discussion	130
5	Discussion and Future Research	133
5.1	Extensions of Variable Selection for Progressive Processes	133
5.2	Addressing Use of Treatment	135
5.3	Extensions of Variable Selection for Illness-death Models	136
5.4	Model Selection Criteria for Interval-censored Data	138
5.5	High-dimensional Problems with $p > n$	140
5.6	Further Considerations in Model Specification and Predictive Assessment	140
	Bibliography	143
	APPENDICES	160
A	Appendix to Chapter 2	161
A.1	Simulation Results with A Smaller Sample Size	161
A.2	Multistate Event Processes under Dependent Visit Scheme	162

B	Appendix to Chapter 3	171
B.1	Computation of Conditional Probabilities	171
B.2	Data Dictionary for Candidate Variables	175
B.3	Variable Selection Performance for Right-censored Data	175
C	Appendix to Chapter 4	180
C.1	Derivations for Expectations Involved for Calculating Limiting Values for Cox Regression.	180
C.1.1	Cox Regression in a Three-state Markov Process	181
C.1.2	Cox Regression in a Semi-Markov Process	183
C.2	Derivations for Calculating Limiting Value for 3-state Model	185
C.2.1	3-state Markov Illness-death model in a Markov Process	185
C.2.2	Markov Illness-death Model in a Three-state Semi-Markov Process .	186
C.2.3	State Occupancy Probabilities in a Semi-Markov Process	188

List of Figures

1.1	Two graphical examples of multistate models	3
2.1	(a) Estimated cumulative distribution functions for the times between radiological assessments based on a semi-Markov model with a gamma frailty and (b) the censoring regions for time to unilateral damage and axial disease for 20 selected patients.	18
2.2	A progressive multistate process with a healthy state (state 0) and k diseased states (state 1 to K)	19
2.3	Turnbull (Turnbull, 1976) estimates with pointwise 95% confidence intervals the cumulative distribution function of the time to entry to state 1 (left panel) and state 2 (right panel)	41
2.4	(a) Changes in the sample adjusted BIC (SABIC) and (b) shrinkage estimates of coefficients from penalized regression with the LASSO penalty across values of $\log \gamma$	44

3.1	(a) Kaplan–Meier estimates of survival probabilities for overall survival and progression-free survival (PFS) with 95% confidence bands; (b) sample paths for four selected subjects with disease status at the last visit, 1 indicating cognitive impairment and 0 being healthy, and death or censoring at the end of study.	53
3.2	A 3-state illness-death model characterizing the onset of cognitive impairment for the samples from the NACC	54
3.3	(a) Changes in SABIC with respect to $\log(\gamma_{\text{LASSO}})$ using LASSO penalty and (b) heatmap of SABIC with respect to two tuning parameters $\log \gamma_{01}$ and $\log \gamma_{\text{death}}$ for LASSO and GLASSO, respectively	79
3.4	Shrinkage estimates of coefficients from penalized EM algorithm with a single LASSO penalty across values of $\log \gamma$	80
4.1	Two representations of the illness-death process.	92
4.2	Limiting values for the regression coefficient estimator under Cox regression involving a binary covariate $X_1 \sim \text{Bernoulli}(0.5)$. Panels (a)–(b): Settings A and B, varying $r = \lambda_{12}/\lambda_{02}$; (c)–(d): Settings 7–8, fix $r = \lambda_{12}/\lambda_{02} = 1, 2$ or 4, varying over β_{01} ; (e)–(f): Settings 9–10, varying $d = \beta_{12}/\beta_{02}$	95

4.3	Limiting values of the estimators for regression coefficients β_{01} , β_{02} , and β_{12} under the Markov illness-death model and of ψ under Cox regression plotted as functions of the common regression coefficient for X_2 on all three transitions when X_2 is omitted from the model. Note X_1 has common regression coefficient of $\log 1.5$ across all three transitions and $r = \lambda_{12}/\lambda_{02} = 1$. In each panel, the horizontal red dashed line indicates the true effect of X_1 , while the black, red, and green curves correspond to limiting values ψ_{01}^* (or ψ_{02}^*), ψ_{12}^* , and ψ^* , respectively.	100
4.4	Limiting values of regression coefficient estimator of β_{12} under Markov illness-death model and ψ under a Cox regression involving a single binary covariate $X_1 \sim \text{Bernoulli}(0.5)$, varying over κ_{12} in Settings 11 and 12.	102
4.5	Simplified multistate diagrams considered for the purpose of predicting specific marginal features	104
4.6	Limiting values of estimated survival functions by X_1 for Cox regression (in black) and three-state model (in red) in Settings 1–6 when $r = \lambda_{12}/\lambda_{02} = 2$	110
4.7	Expected Brier Score evaluated at the limiting value of estimated survival functions derived from a fitted Cox regression (in black) and three-state model (in red) in Settings 1–6 when $r = \lambda_{12}/\lambda_{02} = 2$	111
4.8	Expected Brier Score evaluated at the limiting value of estimated survival functions derived from a fitted Cox regression (in black) and three-state model (in red) in Settings M1, M2, and M3 when $r = \lambda_{12}/\lambda_{02} = 2$	113

4.9	Expected Brier Score evaluated at the limiting value of survival functions derived from the true model (black dotted line), the fitted Markov illness-death model (red dashed line), and the fitted Cox regression (green solid line) when X_2 is omitted from the model. X_1 has homogeneous effect of $\log 1.5$ across all three transitions, X_2 has homogeneous effect of $\log 0.25$ across all three transitions, and $r = \lambda_{12}/\lambda_{02} = 1$. The three panels correspond to the odds ratio between X_1 and X_2 with $OR = 0.5, 1$ and 2	114
4.10	Expected Brier Score derived from the true model (in black), the fitted Markov illness-death model (in red), and the fitted Cox regression (in green) when the true process follows an illness-death semi-Markov model. The two panels correspond to Setting 1 and Setting 2, respectively, with $\kappa_{12} = 0.5$ and $r = \lambda_{12}/\lambda_{02} = 1$	115
4.11	Cumulative incidence functions $F(t)$ in Setting C1, C2, and C3 with $r = \lambda_{12}/\lambda_{02} = 1$, and $p_1 = E_X[P(T_{01} < T_{02} \mid X)] = 0.3, 0.5$ and 0.7 . Black curves represent Fine-Gray estimates, and red curves represent true Markov illness-death model	121
4.12	Limiting value of the expected Brier Score $BS(t)$ in Settings C1, C2, and C3 and $p_1 = E_X[P(T_{01} < T_{02} \mid X)] = 0.3, 0.5$, and 0.7 . Solid curves represent Fine-Gray estimates, and dashed curves represent true Markov illness-death model	122
4.13	Empirical Brier score $\widehat{BS}(t)$ based on the estimated survival functions derived from the Markov illness-death model and Cox regression for overall survival, evaluated on an independent test set of 200 subjects; $r = \lambda_{12}/\lambda_{02} = 1.2$	126

4.14	Empirical Brier score $\widehat{BS}(t)$ based on the estimated cumulative incidence functions derived from the Markov illness-death model and Fine-Gray model for the intermediate (non-fatal) event evaluated based on an independent test set of 200 subjects; $r = \lambda_{12}/\lambda_{02} = 1.2$	126
4.15	a) Brier score for predicting overall survival comparing Cox model and 3-state model, with hashmarks on the top representing the 20%, 40%, 60%, 80% quantiles of death times; b) Brier score for predicting disease progression survival comparing Fine-Gray model and 3-state model, with hashmarks on the top representing the 20%, 40%, 60% quantiles of time to fractures . . .	129
5.1	(a) A more general progressive disease process for modeling the onset of comorbidities and death and (b) A state space diagram for a joint model for a 3-state progressive process $\{Z(t), t \geq 0\}$ and a visit process $\{A(t), t \geq 0\}$	134
5.2	A diagram for a unidirectional progressive process with $K + 1$ states	138
A.1	A state space diagram for a joint model for a 3-state progressive event $Z(t)$ and a visit process $A(t)$	163

List of Tables

2.1	Empirical results for multistate state analysis under independent intermittent observation scheme summarizing transition-specific and covariate-wise true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression following ad hoc right-endpoint or mid-point imputation; $n = 1000$ with $\rho = 16$ as the mean number of visits over $(0, 1]$.	37
2.2	Empirical results for multistate state analysis under independent intermittent observation scheme summarizing transition-specific and covariate-wise true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression following ad hoc right-endpoint or mid-point imputation; $n = 1000$ with $\rho = 4$ as the mean number of visits over $(0, 1]$.	38
2.3	Selected HLA markers for the onset and progression of SI joint damage including unilateral SI joint damage (state 1) and axial disease (state 2) and their effects by multistate penalized EM algorithm and penalized Cox regression after right-endpoint and mid-point imputation using a LASSO penalty	43
3.1	Parameter values used for Settings 1, 2, and 3 in simulation studies	68

3.2	Empirical results for multistate analysis under independent intermittent observation scheme summarizing true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression for PFS and OS time; $n = 1000$ with $\rho = 16$ as the mean number of visits over $(0, 1]$	74
3.3	Empirical results for multistate analysis under independent intermittent observation scheme summarizing true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression for PFS and OS time; $n = 1000$ with $\rho = 8$ as the mean number of visits over $(0, 1]$	75
3.4	Selected covariates and estimated coefficients by the 3-state model and the Cox model	81
4.1	Coefficients for five settings involving a single binary covariate $X_1 \sim \text{Bernoulli}(0.5)$ with constraints $\lambda_{12}/\lambda_{02} = r$	94
4.2	Covariate coefficients (β) for different transitions across settings and five independent covariates x_1 – x_5 generated from Bernoulli (0.5). $r = \lambda_{12}/\lambda_{02} = 2$	112
4.3	Coefficients for three settings involving a single binary covariate X_1 , $\lambda_{12}/\lambda_{02} = r = 1$, and $X_1 \sim \text{Bernoulli}(0.5)$	119
4.4	Estimated coefficients (EST), empirical standard errors (ESE), and robust standard errors (RSE) when data are generated under the 3-state model with $r = \lambda_{12}/\lambda_{02} = 1.2$, and analyzed using the 3-state Markov model, Cox model, and Fine-Gray model.	125

4.5	Summary of covariate effects (EST), robust standard errors (RSE), and p-values for the 3-state Markov model, Cox model for overall survival, and Fine-Gray model for fracture.	128
A.1	Empirical results for multistate state analysis under independent intermittent observation scheme summarizing transition-specific and covariate-wise true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression following ad hoc right-endpoint or mid-point imputation; $n = 500$ with $\rho = 16$ as the mean number of visits over $(0, 1]$. .	166
A.2	Empirical results for multistate state analysis under independent intermittent observation scheme summarizing transition-specific and covariate-wise true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression following ad hoc right-endpoint or mid-point imputation; $n = 500$ with $\rho = 4$ as the mean number of visits over $(0, 1]$. .	167
A.3	Empirical results for multistate state analysis under dependent intermittent observation scheme summarizing transition-specific and covariate-wise true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression following ad hoc right-endpoint or mid-point imputation; $n = 1000$ with $\rho = 4$ as the mean number of visits over $(0, 1]$ under Setting 1.	168

A.4	Empirical results for multistate state analysis under dependent intermittent observation scheme summarizing transition-specific and covariate-wise true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression following ad hoc right-endpoint or mid-point imputation; $n = 1000$ with $\rho = 4$ as the mean number of visits over $(0, 1]$ under Setting 2.	169
A.5	Empirical results for multistate state analysis under dependent intermittent observation scheme summarizing transition-specific and covariate-wise true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression following ad hoc right-endpoint or mid-point imputation; $n = 1000$ with $\rho = 4$ as the mean number of visits over $(0, 1]$ under Setting 3.	170
B.1	Description and preprocessing of selected candidate variables	174
B.2	True and false positive rates (%) for all methods across three settings. Penalized multistate results are reported with both non-parametric and piecewise-constant baselines; Cox-based models include OS and PFS settings.	177
B.3	True and false positive rates (%) for the LASSO-based models across three settings.	178

Chapter 1

Introduction

1.1 Overview

In medical studies, patients experience events over time and given the process history and demographic information, interest often lies in their prediction, or simply prediction of the disease status at some time horizon. Prediction may be viewed as making use of observed data to “estimate” future observations, and it has played an important role in disease prevention and management in many contexts of health research ([Steyerberg, 2009](#)). Variable selection algorithms are particularly useful for improving predictive accuracy by trading an increase in bias for a larger reduction in variance, thereby enhancing the model’s overall predictive performance ([Hastie and others, 2009](#)). They are also useful for identifying a smaller subset of predictors which has the strongest effects on the disease processes from a large number of variables.

In this thesis, we are concerned with use of statistical models and methods for time-to-event data and disease processes to develop innovative methods for the identification of

important covariates in a disease process and to make good predictions about the future disease course. The thesis discusses three main topics including

- (i) Variable selection for the progression of chronic diseases with high-dimensional covariates under intermittent observation,
- (ii) Variable selection for progressive disease processes with an absorbing state (typically death) under dual censoring scheme,
- (iii) Assessing the implications of model misspecification on regression coefficients estimation and prediction in multistate analysis.

This introductory chapter includes a brief overview of statistical models and methods used in this thesis, introduces scientific research programs motivating the studies, and outlines the subsequent research projects.

1.2 Statistical Models and Methods for Lifetime Data

This section gives a brief overview of several statistical tools which our research is based on in this thesis. Chapters 2, 3, and 4 are written to be self-contained to ensure clarity and readability so there is some repetition in the presentation of notation and specification of foundational functions for modeling.

Multistate models are particularly useful to analyze life history data and we consider them here for studying disease processes. In multistate models, distinct disease states are identified to define a state space and arrows in state space diagram depict which transitions are possible at any instant in time. Events of interest could be considered as transitions

from one state to another. Interests may lie in transition times, the risk of making certain types of transitions, state occupancy probabilities, the effects of external covariates, and so on. [Cook and Lawless \(2018\)](#) discussed various features of life history data and illustrated how to use suitable multistate models to analyze the data according to its nature. We introduce basic notation and the formulation of multistate models in Section [1.2.1](#).

1.2.1 Multistate Processes

Intensity functions play a central role in characterizing multistate processes by defining the conditional instantaneous probability an individual moves into a new state given their current state and the disease history. Figure [1.1](#) shows two graphical examples of multistate models.

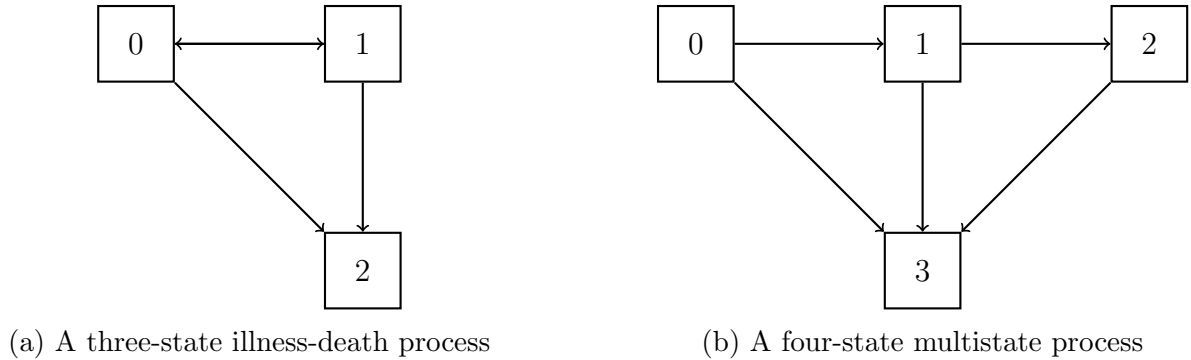


Figure 1.1: Two graphical examples of multistate models

Let $\mathcal{S} = \{0, 1, \dots, K\}$ be a state space and $Z(t)$ denote the state occupied at time t for a generic individual. Let $H(t) = \{Z(u), 0 \leq u < t\}$ be the history up to time t which contains information about state occupancy during the interval $[0, t)$; here we consider a simple case without any covariates. The intensity function for a transition from state k to

state l is defined as

$$\lambda_{kl}(t \mid H(t)) = \lim_{\Delta t \rightarrow 0} \frac{P(Z(t + \Delta t^-) = l \mid Z(t^-) = k, H(t))}{\Delta t}, \quad (1.1)$$

for $k, l = 0, 1, 2, \dots, K$ and $k \neq l$. The subscript k in $\lambda_{kl}(t \mid H(t))$ reflects the fact that this is defined given $Z(t^-) = k$.

It is common to adopt a Markov assumption in which case the time scale is the age of the process and the intensity only depends on the current state occupied rather than the history. In Markov models, the intensity function can be simplified as

$$\lambda_{kl}(t \mid H(t)) = \lambda_{kl}(t). \quad (1.2)$$

Let $\Lambda_{kl}(t) = \int_0^t \lambda_{kl}(s) ds$ denote a cumulative intensity and $\mathbb{Q}(t)$ denote a $(K+1) \times (K+1)$ matrix of transition intensities, where the (k, l) entry is $\lambda_{kl}(t)$ for $k \neq l$ and $-\sum_{l \neq k} \lambda_{kl}(t)$ is in the k th diagonal entry. We denote the $(K+1) \times (K+1)$ transition probability matrix as $\mathbb{P}(s, t)$ in which the (k, l) entry is the probability that state l is occupied at time t given that state k was occupied at time s ; $[\mathbb{P}(s, t)]_{kl} = P(Z(t) = l \mid Z(s) = k)$. Note that $\mathbb{P}(s, t)$ can be expressed via a product integral

$$\mathbb{P}(s, t) = \prod_{u \in [s, t)} \{\mathbb{I} + \mathbb{Q}(u) du\},$$

where $\mathbb{I}$ is a $(K+1) \times (K+1)$ identity matrix. For time-homogeneous Markov models, let $\mathbb{Q}$ be the matrix of transition intensities. $\mathbb{P}(s, t)$ can be straightforwardly expressed as

$$\mathbb{P}(s, t) = \exp(\mathbb{Q}(t - s)), \quad (1.3)$$

where the exponential function refers to matrix exponential.

1.2.2 Multistate Processes under Complex Observation Scheme

In some observational studies, the responses can only be evaluated by laboratory tests or radiological examination and subjects are only intermittently assessed at periodic visits. For example, in a study on the effectiveness of zoledronic acid on skeletal metastases, the status of bone disease was determined through periodic bone scans at 3, 6, and 9 months ([Rosen and others, 2003](#)); retinopathy of diabetes patients was measured using fundus photography and graded on an ordinal scale every 6 months in the Diabetes Control and Complications Trial (DCCT) ([Nathan and Group, 2014](#)). Under these situations, the states characterizing the stage of the disease are only known at the assessment times, so that the exact transition times and the number of transitions are unknown, resulting in incomplete data under intermittent observation. In some scenarios such as cancer studies, the disease progression status is only observed at intermittent assessment times, resulting in interval-censored progression time, while the time to death is subject to right-censoring. For individuals who are progression-free at their last assessment, their ultimate progression status is unknown since they may or may not have progressed after the final visit. The setting in which progression times are interval-censored and death times are subject to random right-censoring is referred to as dual observation scheme ([Boruvka and Cook, 2016](#)).

There has been much work on model fitting for multistate models under intermittent observation. [Kalbfleisch and Lawless \(1985\)](#) developed a Fisher-scoring algorithm for estimation and inference under the Markov assumption, and the `msm` package ([Jackson, 2011](#)) in R can be used for fitting Markov models while accommodating homogeneous or piece-

wise constant baseline intensities. Such analyses are justified under the assumption of the conditionally independent visit process (CVIP) assumption [Cook and Lawless \(2021\)](#). Semi-Markov models are more challenging due to the intractable likelihood with interval-censored event times. Therefore, early work on model fitting for semi-Markov models under intermittent observation is restricted to special cases such as progressive processes ([Satten and Sternberg, 1999](#); [Sternberg and Satten, 1999](#)). [Foucher *and others* \(2007\)](#) considered more general progressive semi-Markov models with multiple terminal events and proposed a parametric approach to estimate the hazard functions. [Kang and Lagakos \(2007\)](#) allow reversible transitions but restrict the intensity function for at least one of the states to be time-homogeneous. A more flexible framework for fitting semi-Markov interval-censored data was proposed by [Titman and Sharples \(2010\)](#) which makes use of latent-state continuous time Markov chain to accommodate non-exponential sojourn time distributions and more general reversible processes. [Lange and Minin \(2013\)](#) described an EM algorithm for maximizing the likelihood in [Titman and Sharples \(2010\)](#).

1.2.3 Variable Selection for Time-to-event Analysis

Variable selection algorithms offer a useful approach for identifying prognostic variables and building predictive models and such methods have been adapted to deal with failure times and some other dynamic processes. Through shrinkage of regression coefficients, ridge regression enhances predictive accuracy by mitigating over-fitting, but it does not shrink coefficients to zero and so cannot be used for variable selection. The “least absolute shrinkage and selection operator” (LASSO) was proposed by [Tibshirani \(1996\)](#) to achieve this. Selection consistency ([Zou, 2006](#)) is a desirable property, in which the probability of selecting the correct set of variables approaches one as the sample size goes to infinity –

the LASSO does not have this property. Through the incorporation of weights, the *adaptive* LASSO penalty (Zou, 2006) ensures selection consistency and moreover the estimators have the same asymptotic distribution as maximum likelihood estimators from the model containing the correct set of covariates. When covariates (e.g., markers) are grouped in terms of their functions and interest lies in selecting sets of variables together, the *group* LASSO can be used. Yuan and Lin (2006) proposed such a penalty term for linear regression and Meier and others (2008) extended the approach for generalized linear modeling. Wang and others (2007) introduced the *adaptive group* LASSO which ensures selection consistency in this setting.

Most of the recent work on variable selection for failure time data has focused on right-censored observations and uses the Cox model with its associated partial likelihood (Tibshirani, 1997; Simon and others, 2011). This can now be implemented in the `glmnet` package in R (R Core Team and others, 2013). For interval-censored failure times, Wu and Cook (2015) adopted piecewise-constant baseline hazards in a proportional hazards model and defined a penalized observed data likelihood. Zhao and others (2020) introduced a broken adaptive ridge (BAR) regression for variable selection in interval-censored failure time data. This approach has been generalized to accommodate group variable selection in the Cox model with interval-censored failure time data, and methods for broader classes of survival models and mixture models have also been proposed; see Li and others (2020), Tian and Sun (2023), and Xu and others (2021). In the context of multistate models for right-censored data, Sennhenn-Reulen and Kneib (2016) applied the LASSO and fused LASSO penalties to select covariates and to smooth covariate effects across different transitions.

1.3 Motivating Research Programs

The motivation for the methods considered in this thesis came from three different scientific research programs, including (1) a longitudinal cohort maintained by the University of Toronto Psoriatic Arthritis Clinic (UTPAC), where patients with psoriasis were followed and underwent periodic examinations to assess joint damage; (2) the National Alzheimer’s Coordinating Center (NACC), which harmonized longitudinal data from 37 Alzheimer’s Disease Research Centers (ADRCs) across the United States; and (3) an international, multi-center randomized trial conducted by Novartis Pharmaceuticals involving women with stage IV breast cancer metastatic to bone. These are described briefly in what follows.

1.3.1 The University of Toronto Psoriatic Arthritis Clinic (UT-PAC)

The University of Toronto Psoriatic Arthritis Clinic (UTPAC) was founded in 1976 and has since maintained a registry of over 1600 psoriatic arthritis (PsA) patients. Patients are recruited to this tertiary care center by a variety of means including physician referral and screening of high risk individuals. Upon recruitment to the registry patients provide demographic information, provide a medical history, undergo clinical examination and provide blood samples for genetic testing. A baseline radiographic examination is carried out, and annual clinic and biannual radiographic examinations are scheduled to assess joint damage according to a standardized protocol ([Gladman and Chandran, 2011](#)). Although assessments are scheduled at regular intervals, patients tend to visit the clinic irregularly, which leads to considerable variation in both the frequency and timing of visits across individuals. The extent of joint damage was assessed by clinicians in clinic assessments

and graded based on a different five-point scale in radiological evaluations ([Rahman and others, 1998](#)). Overall, the UTPAC cohort provides a rich data set for researchers to study the course of this autoimmune disease, identify risk factors for rapid disease progression, and to evaluate patient responses to treatment.

1.3.2 The National Alzheimer’s Coordinating Center (NACC)

The National Alzheimer’s Coordinating Center (NACC) maintains a data repository that collects longitudinal data from 37 Alzheimer’s Disease and Research Centers (ADRCs) across the United States. We utilized the Uniform Data Set (UDS), which has been collected using a standard evaluation protocol and includes all subjects enrolled across ADRCs since 2005, with data frozen as of March 2024. Each ADRC has its own recruitment protocol according to center-specific research goals, but the specific ADRC associated with each participant is not identifiable in the UDS. Subjects are typically enrolled through a combination of clinical referral, self-referral, and active recruitment, and many centers also recruit cognitively normal individuals. The NACC UDS cohort is not representative of the general population in the United States and therefore, is not appropriate for estimating the prevalence or incidence of cognitive impairment or dementia. See <https://naccdata.org/requesting-data/nacc-data> for a description of the UDS study population and for data request, and [Arce Rentería and others \(2023\)](#) for a discussion on the representativeness of the NACC cohort.

1.3.3 A Randomized Trial of Patients with Skeletal Metastases

The data is from an international double-blind phase III randomized trial for patients with breast cancer metastatic to bone ([Hortobagyi and others, 1996](#)). The primary goal for this

trial is to evaluate the effectiveness of pamidronate in preventing skeletal complications, including fractures, the need for radiation to bone or bone surgery, and spinal cord compression. Women with stage IV breast cancer who had no skeletal complications during the two weeks prior to enrollment were recruited between January 1991 and March 1994 from 97 study sites across the United States, Canada, Australia, and New Zealand. Among the 380 women recruited, 185 of them were randomly assigned to receive monthly infusions of pamidronate, and 195 to receive placebo. Participants were scheduled for monthly visits and were assessed for the occurrence of skeletal complications at each visit. They are followed until death, loss to follow-up, or the end of study on February 1, 1995.

1.4 Introduction to Research Topics

1.4.1 Variable Selection for Progressive Multistate Models

As PsA is a chronic condition that can cause joint damage and progressively worsen, the process of joint damage can be described by progressive multistate models. Research has found that human leukocyte antigen (HLA) antigens play an important role in the disease progression of psoriatic arthritis ([Gladman *and others*, 1998](#)). Among a large group of HLA candidates, we aim to identify the important ones associated with the progression of joint damage in PsA. Since patients tend to attend clinic visits at irregular intervals, and joint damage can only be assessed through clinical or radiological examinations conducted during these visits, the development of joint damage is not observed continuously, but instead intermittently. Chapter 2 discusses variable selection in progressive multistate models when disease processes are under intermittent observation. We describe an Expectation-Maximization (EM) algorithm which can flexibly incorporate different penalty functions

for variable selection and allows one to exploit existing packages for penalized regression in the maximization step. Simulation studies are conducted to empirically show the performance of identifying important effects with different penalty functions. Application of the proposed algorithm for different penalty functions to data from the UTPAC clinic involves selecting HLA markers associated with joint damage progression in PsA patients.

1.4.2 Variable Selection for Illness-death Processes under Dual Observation Schemes

In many cases, subjects may follow a disease progression involving multiple stages that can be described using a multistate model, while also being at risk of death. To account for this, it is ideal to jointly model both the disease progression and the death process. This approach is particularly useful in studies where death poses a non-ignorable competing or semi-competing risk, such as in cancer research or studies involving elderly populations. A commonly used framework in such settings is the classical illness-death model which offers a useful framework for studying the progression of chronic disease while jointly modeling death. In many settings, the time of disease progression is not observed directly, but progression status is recorded at intermittent assessment times. This creates a dual observation scheme where progression times are interval-censored and death is subject to right censoring. In Chapter 3, we present a penalized observed-data likelihood for variable selection in multiplicative intensity-based models for the joint process, involving different penalty functions for different sets of regression coefficients. Optimization is carried out based on an innovative expectation-maximization algorithm that can be implemented using existing software. Simulation studies demonstrate the finite sample performance of the method, and an application to a dementia study from the National Alzheimer’s Coordi-

nating Center (NACC) ([Beekly and others, 2007](#)) is provided to illustrate the insights that can be gained from this approach.

1.4.3 Model Misspecification in Multistate Analysis and Implications on Prediction

Illness-death models provide a useful framework for jointly analyzing disease progression and survival. However, simpler approaches are routinely used such as the Cox proportional hazards model for overall survival and the Fine-Gray model for competing risk analysis. Which framework is best for any given problem is often not clear. Multistate models offer a comprehensive framework analysis and prediction, but require stronger assumptions and a greater degree of specification. When interest lies in a particular feature, development of models geared towards that feature (e.g., survival) are more direct and simple. Investigations of the strengths and weaknesses of competing methods is therefore worthwhile.

In Chapter 4, we examine the limiting values of regression estimators from Cox and Fine-Gray models when the underlying data-generating process follows an illness-death model. We further assess the impact of model misspecification on predictive performance as well as robustness against omission of important covariates. An application to predicting overall survival time and fractures among patients with metastatic breast cancer is provided for illustration.

Chapter 2

Variable Selection for Progressive Multistate Processes Under Intermittent Observation

2.1 Introduction

2.1.1 Background

Multistate models offer a powerful framework for understanding chronic disease processes through estimation of progression rates, examination of covariate effects, and for prediction of the disease course ([Cook and Lawless, 2018](#); [Andersen and Ravn, 2023](#)). Through use of established methods, important insights have been gained about chronic disease processes in hepatitis ([Sweeting and others, 2010](#)), arthritis ([Cook and Lawless, 2014](#); [Sutradhar and Cook, 2008](#)), neurological conditions ([Shi and others, 2021](#); [Perera and others, 2023](#)), and

other settings. When it is only possible to determine state occupancy based on laboratory tests (Lawless and Nazeri Rad, 2015), at clinic visits (Sutradhar and Cook, 2009), or upon radiological examination (Cook and Lawless, 2014), transition times between states are unknown. In a highly influential paper, Kalbfleisch and Lawless (1985) developed a Fisher-scoring algorithm for fitting Markov models when such processes are under intermittent observation. This algorithm is implemented in the `msm` package (Jackson, 2011) in R and accommodates intensity-based regression modeling. Such analyses are justified under the assumption of a conditionally independent visit process (CVIP) as described in (Cook and Lawless, 2021).

There has been relatively little work on the use of penalty functions for variable selection in the context of multistate analyses. Sennhenn-Reulen and Kneib (2016) proposed use of the LASSO and fused LASSO (Tibshirani *and others*, 2005) algorithms for reducing the dimension of parameters indexing transition intensities for multistate processes under right-censoring where the latter penalty term was used to constrain some covariate effects to be the same for different transitions; see also Dang *and others* (2021). For processes under intermittent observation, Wu and Cook (2015) considered the special case of interval-censored failure time data and examined a variety of penalty functions for variable selection based on a proportional hazards formulation. Zhao *and others* (2020) proposed a broken adaptive ridge (BAR) regression for variable selection in this same context. Li *and others* (2020) considered penalized methods for variable selection based on general transformation models for failure time analysis. Sun *and others* (2022) addressed the problem for multivariate interval-censored data. In the multistate setting, Machado *and others* (2021) used penalization to fit smoothing splines for estimation of transition intensities when processes are under intermittent observation. Motivated by a collaboration with clinician scientists studying risk factors for joint damage in arthritis, here we develop and

evaluate penalized likelihood for identifying markers associated with disease progression when a multistate process is under intermittent observation. We focus on progressive processes and consider a complete data likelihood in which the transition times are known and formulate an expectation-maximization ([Dempster *and others*, 1977](#)) algorithm which enables use of established software for variable selection. The formulation we adopt enables use of existing software for variable selection for generalized linear models.

The remainder of the chapter is organized as follows. In the following sub-section, we briefly discuss the University of Toronto Psoriatic Arthritis Registry where the scientific agenda motivates our work. In [Section 2.2](#) we introduce notation and develop the observed data likelihood for Markov processes under intermittent observation. A penalized log-likelihood is then introduced before we recast the problem in terms of a penalized complete data likelihood in [Section 2.3](#) which is to be maximized via an EM algorithm for variable selection. In [Section 2.4](#), we conduct simulation studies to evaluate the finite sample performance of the proposed algorithm with the LASSO, adaptive LASSO, group LASSO, and adaptive group LASSO penalties, and compare these methods to some ad hoc methods based on failure time data. We propose novel metrics (true and false positive rate) to judge the relative performance of the various penalty functions in the multistate setting. We also examine the robustness of the proposed algorithm to a conditionally dependent visit process. [Section 2.5](#) includes an illustrative application of the proposed algorithm to model the risk of sacroiliac (SI) joint damage and assess the role of HLA markers. In [Section 2.6](#), we make some concluding remarks and discuss areas warranting future research.

2.1.2 HLA Markers and Sacroiliac Joint Damage in PsA

A central aim is to assess the role of genetic markers on the rate of disease progression in order to identify high risk individuals who may warrant more careful monitoring or powerful but costly prophylactic and therapeutic interventions. We consider registry data from the UTPAC comprised of 1630 individuals among which 1006 were typed for their human leukocyte antigen (HLA) profile including those in the HLA-A, HLA-B, HLA-C, HLA-DR and HLA-DQB regions. We restrict attention to HLA-B and HLA-C markers as these are the regions of primary interest and after excluding markers with frequencies less than 1%, 41 HLA candidate markers remain for consideration.

A feature of particular interest is the extent of damage of the SI joints since back involvement leads to serious pain and impaired functional ability. The degree of damage of the sacroiliac joints is assessed based on the modified New York criterion ([Linden and others, 1984](#)) using plain radiographs of the pelvis; each joint is graded so that a score of 0 is given for a normal SI, 1 if it has suspicious changes, 2 if there is minimal abnormality, 3 if there is unequivocal abnormality, and the highest score of 4 denotes severe abnormality (e.g. complete joint destruction). These scores are available for the left and right SI joints and we use these scores to characterize the development of SI joint damage using a three-state progressive model including an initial disease-free state (state 0), unilateral SI joint damage (state 1) defined as having a severity score of 2 on *at most one side*, and axial disease (state 2), defined as having a score of at least 2 on *both sides*, or having a score of 3 or more on one side. Thus, the transition time from state 0 to 1, denoted T_1 , represents the onset of unilateral SI joint damage and the transition time from state 1 to 2 (T_2 where $T_2 > T_1$) represents the progression time from unilateral SI joint damage (on either side) to axial disease. While radiographic examinations are scheduled every two years, there is

considerable heterogeneity between individuals and within individuals over time in their propensity to undergo X-rays making the examination times highly variable. Figure 2.1 (a) displays the estimated cumulative distribution functions for the times between successive radiographic examination times for the first ten assessments. The estimates are obtained by fitting a conditionally semi-Markov (gap time) model with a patient-level gamma frailty introduced to address the within-patient dependence in successive gap times, and stratified by the visit number. The realized visit times for each individual (along with their transition times) determine the censoring regions for (T_1, T_2) . Specifically the interval $\mathcal{C}_1 = (L_1, R_1)$ containing the T_1 transition time is defined, with L_1 the last assessment time an individual was known to be in state 0 and R_1 the first assessment time they were known to be in state 1. Likewise, $\mathcal{C}_2 = (L_2, R_2)$ contains the T_2 transition time with L_2 the last radiographic assessment time an individual was known not to have entered state 2 and R_2 the first assessment time they were known to be in state 2. The joint censoring regions are defined by the Cartesian product $\mathcal{C} = \mathcal{C}_1 \times \mathcal{C}_2$. Figure 2.1 (b) shows a sample of 20 censoring regions for the times to unilateral damage (state 1) and axial disease (state 2) for 20 patients where time is measured from clinic entry. The heterogeneity in the size and shape of censoring regions shows the variation in the information about the entry times across patients.

We aim to select human leukocyte antigens (HLA) which are associated with the progression of SI joint damage from a large number of candidates, based on the 3-state model while accounting for the fact that individuals are under intermittent observation.

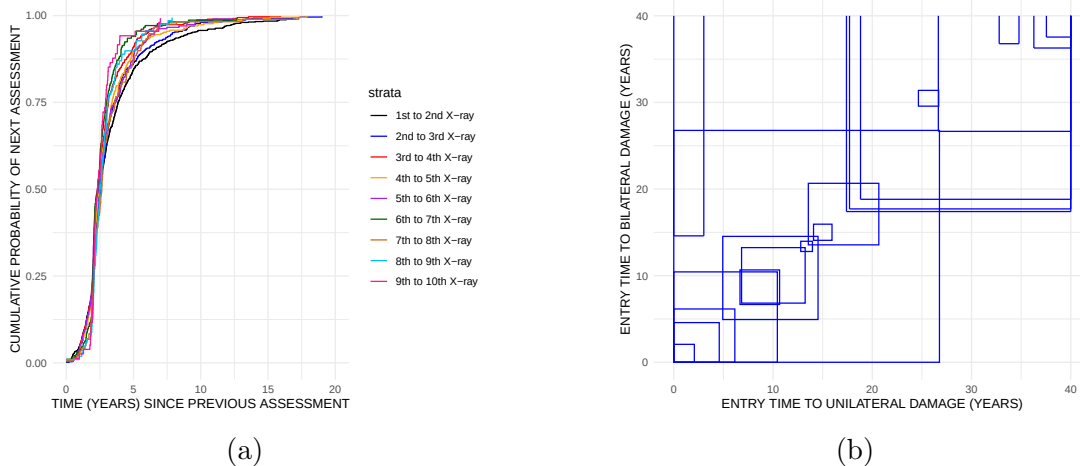


Figure 2.1: (a) Estimated cumulative distribution functions for the times between radiological assessments based on a semi-Markov model with a gamma frailty and (b) the censoring regions for time to unilateral damage and axial disease for 20 selected patients.

2.2 Variable Selection for Transition Intensities

2.2.1 Multistate Model and Likelihood Construction

We consider progressive disease processes which can be modeled by a $(K+1)$ -state process as depicted in Figure 2.2 where states are labeled such that those with higher numbers correspond to more advanced stages of damage. Here state 0 represents the initial state of being damage-free, while states $1, \dots, K-1$ represent stages of increasing levels of damage and K is occupied when the disease has fully progressed. We consider a sample of n independent processes and let i index individuals, $i = 1, \dots, n$. Let $Z_i(t)$ denote the state occupied at time t by individual i with $Z_i(0) = 0$ and T_{ik} denote the transition time from state $k-1$ to state k for individual i , $k = 1, 2, \dots, K$. We let $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$ be a $p \times 1$ covariate vector and $H_i(t) = \{Z_i(t), 0 \leq s \leq t, \mathbf{x}_i\}$ the disease history; we restrict attention to the setting of fixed covariates. Consider a conditionally Markov model for which the

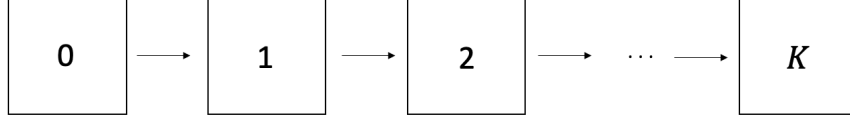


Figure 2.2: A progressive multistate process with a healthy state (state 0) and k diseased states (state 1 to K)

intensity function governing state $k - 1$ to k transitions is

$$\lambda_k(t \mid \mathbf{x}_i) = \lim_{\Delta t \rightarrow 0} \frac{P(Z_i(t + \Delta t^-) = k + 1 \mid Z_i(t^-) = k, \mathbf{x}_i)}{\Delta t}, \quad (2.1)$$

where t^- is an instant before t . We consider models where covariates act multiplicatively on the transition intensities and take the form

$$\lambda_k(t \mid \mathbf{x}_i) = \lambda_{k0}(t) \exp(\mathbf{x}_i' \boldsymbol{\beta}_k),$$

where $\lambda_{k0}(t)$ is the baseline intensity if $\mathbf{x}_i = \mathbf{0}$, and $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \dots, \beta_{kp})'$ is a $p \times 1$ vector of regression coefficients associated with the transition from $k - 1$ to k , $k = 1, \dots, K$.

We consider the scenario in which the state individuals occupy can only be determined at periodic assessment times. We let $A_{i0} = 0$ and $A_{i1} < \dots < A_{iM_i}$ denote the random assessment times for individual i over $(0, \tau_i)$ where τ_i is an administrative censoring time. The realized number of visits is m_i with the realized visit times denoted $a_{i1} < \dots < a_{im_i} < \tau_i$ with $a_{i0} = 0$ with probability one. For a sample of n independent processes, the observed data is the visit times $\{a_{ir}, r = 1, 2, \dots, m_i, i = 1, \dots, n\}$, the states occupied at the visit times, $\{Z_i(a_{ir}), r = 1, \dots, m_i, i = 1, \dots, n\}$ and the values of covariates $\{\mathbf{x}_i, i = 1, \dots, n\}$.

We let $H_i^\circ(a_{ij}) = \{Z_i(a_{ir}), a_{ir}, r = 1, \dots, j, \mathbf{x}_i\}$ denote the history of the observed data up to a_{ij} for individual i , $i = 1, \dots, n$. Conditioning on $H_i^\circ(a_{i0}) = \{Z(a_{i0}), a_{i0}, \mathbf{x}_i\}$, the observed data likelihood for subject i takes the form

$$\prod_{j=1}^{m_i} P(Z_i(a_{ij}), A_{ij} = a_{ij} \mid H^\circ(a_{i,j-1})),$$

which can be factored as

$$\prod_{j=1}^{m_i} P(Z_i(a_{ij}) \mid A_{ij} = a_{ij}, H_i^\circ(a_{i,j-1})) P(A_{ij} = a_{ij} \mid H_i^\circ(a_{i,j-1})). \quad (2.2)$$

When interest lies in the disease process we can drop the term $P(A_{ij} = a_{ij} \mid H_i^\circ(a_{i,j-1}))$ if the visit process is non-informative and if it is leaving the partial likelihood based on n individuals conditionally independent ([Cook and Lawless, 2021](#)) we can treat the assessment times as fixed and, suppressing the notation regarding the assessment times, write the partial likelihood as

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^n \prod_{j=1}^{m_i} P(Z_i(a_{ij}) \mid Z_i(a_{i0}), \dots, Z_i(a_{i,j-1}), \mathbf{x}_i).$$

Under the Markov assumption the observed log-likelihood simplifies further to

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{j=1}^{m_i} \log P(Z_i(a_{ij}) \mid Z_i(a_{i,j-1}), \mathbf{x}_i). \quad (2.3)$$

We next discuss penalization of this log-likelihood with a view to selecting which elements of the covariate vector are important to consider for the state-to-state transitions.

2.2.2 The Penalized Observed Data Likelihood

Variable selection is motivated by the need to improve model interpretability and predictive accuracy and identify prognostic variables, such as genetic markers, which are important to the progression of the disease from a large set of candidates. Recall that β_k is the $p \times 1$ vector of regression coefficients for the intensity of entry to state k , and let $\beta = (\beta'_1, \dots, \beta'_K)'$. The penalized version of (2.3) is

$$\ell_{pen}(\theta) = \frac{1}{n} \ell(\theta) - p_\gamma(\beta), \quad (2.4)$$

where $p_\gamma(\beta)$ denotes the penalty term, and γ is a hyperparameter that controls the extent of the penalty, typically selected through commonly-used model selection criteria such as cross-validation (CV) (Stone, 1974), generalized cross-validation (GCV) (Golub *and others*, 1979), AIC (Akaike, 2003), or BIC (Schwarz, 1978).

Here we consider issues arising in the use of penalty functions in the multistate context. Conventionally penalty functions have a dual role of shrinking coefficients to mitigate over-fit and shrinking some coefficients to zero to aid in variable selection. In most settings different regression coefficients correspond to different covariates so there is a simple relation between coefficients being set to zero and elimination of covariates. In the multistate setting a single covariate can exert several effects if it acts on several transition intensities so penalization can constrain some of covariate effects to zero while allowing others to be non-zero. In its simplest form, the LASSO (Tibshirani, 1996) treats all coefficients as if they are linked to different covariates using the L_1 penalty function of the form $p_\gamma(\beta) = \gamma \sum_{k=1}^K \sum_{j=1}^p |\beta_{kj}|$. In a similar spirit the adaptive LASSO (Zou, 2006) uses adaptive weights giving a penalty term of the form $p_\gamma(\beta) = \gamma \sum_{k=1}^K \sum_{j=1}^p w_{kj} |\beta_{kj}|$, where $w_{kj} = |\tilde{\beta}_{kj}|^{-\nu}$, $\nu \geq 1$, and $\tilde{\beta}_{kj}$ is an estimate of β_{kj} . In practice, $\tilde{\beta}_{kj}$ is often taken to be

the maximum likelihood estimate (MLE) and ν is set to 1 ([Zhang and Lu, 2007](#)).

In the multistate setting it is reasonable for covariates to have a role in multiple intensities and in this sense penalty functions that treat coefficients in groups has some appeal. The rationale is that in progressive processes depicted in [Figure 2.2](#), factors that increase or decrease the rate of transitions across states may often be expected to affect different transitions similarly. This is particularly reasonable when the transitions correspond to similar progression events in an ongoing disease process. Specifically in the setting of a $K + 1$ state model, the j th covariate will give rise to K regression coefficients $\beta_{1j}, \dots, \beta_{Kj}$ and these coefficients can be considered as a group of coefficients. Note that this is slightly different in spirit to the typically application of penalty functions with a “group” feature, but if one considers interactions between $Z_i(t^-)$ and x_{ij} , say, and let $x_{ijk} = I(Z_i(t^-) = k)x_{ij}$ then one can consider the grouping as applying to the derived covariates $x_{ij1}, \dots, x_{ijK}$. The group LASSO can be adapted and the penalty function thereby takes the form

$$p_\gamma(\boldsymbol{\beta}) = \gamma \sum_{j=1}^p \sqrt{\sum_{k=1}^K \beta_{kj}^2}.$$

The adaptive group LASSO penalty is in the same spirit as the adaptive LASSO as it allows for different weights for different groups of coefficients; the penalty term has the form

$$\gamma \sum_{j=1}^p w_j \sqrt{\sum_{k=1}^K \beta_{kj}^2},$$

where one can set $w_j = 1/(\sum_{k=1}^K \tilde{\beta}_{kj}^2)^{1/2}$ with $\tilde{\beta}_{kj}$ is the maximum likelihood estimate ([Wang and Leng, 2008](#)).

We next discuss an expectation-maximization algorithm to maximize the penalized log-

likelihood in (2.4) by considering a complete data likelihood in which the transition times are observed subject to right-censoring.

2.3 Variable Selection via an Expectation-Maximization Algorithm

2.3.1 The Complete Data Log-likelihood

Here we consider a complete data likelihood in which the transition times are taken as part of the complete data. Let $t_{i0} = 0$, t_{ik} denote the realization of the entry time to state k for individual i , and $\lambda_k(t_{ik}|\mathbf{x}_i)$ be the $k-1 \rightarrow k$ intensity function (2.1) evaluated at this time with the corresponding covariate, $k = 1, \dots, K$. Under a Markov model, the complete data likelihood takes the form

$$\mathcal{L} \propto \prod_{i=1}^n \prod_{k=1}^K \lambda_k(t_{ik} | \mathbf{x}_i) \exp(-\Lambda_k(t_{i,k-1}, t_{ik} | \mathbf{x}_i)) ,$$

where $\Lambda_k(s, t|\mathbf{x}_i) = \int_s^t \lambda_k(u|\mathbf{x}_i)du$ is the integrated $k-1 \rightarrow k$ transition intensity for individual i , $i = 1, \dots, n$. This complete data likelihood can be factored as $\mathcal{L} \propto \prod_{k=1}^K \mathcal{L}_k$ where

$$\mathcal{L}_k \propto \prod_{i=1}^n \lambda_k(t_{ik} | \mathbf{x}_i) \exp(-\Lambda_k(t_{i,k-1}, t_{ik} | \mathbf{x}_i)) .$$

We adopt piecewise-constant baseline intensities with $q_k - 1$ cut-points giving q_k parameters for the $k-1 \rightarrow k$ intensity function. Let $0 = b_{k0} < b_{k1} < b_{k2} < \dots < b_{k,q_k-1} < \infty$ denote these cut-points, with $\mathcal{B}_{kh} = (b_{k,h-1}, b_{kh}]$ be the h th interval, $h = 1, \dots, q_k$, where we let b_{kq_k} be infinity, $k = 1, \dots, K$. We then let λ_{kh} be the constant baseline intensity

for $t \in \mathcal{B}_{kh}$, $h = 1, \dots, q_k$, $k = 1, \dots, K$. The number and location of cut-points in the baseline intensities requires specification; [Lawless and Zhan \(1998\)](#) noted that relatively few cut-points are needed to render a reasonably accurate approximation of the underlying intensities, ensuring that estimated regression coefficients have desirable properties and that predictions based on fitted models are reasonably accurate. It is generally desirable for estimated baseline intensity parameters to have comparable standard errors; with right censored data this will often mean choosing cut-points so that there are a comparable number of transitions observed in each interval. When events or transition times are interval-censored nonparametric estimates of relevant distributions can guide selection of cut-points and we use this approach in the application in [Section 2.5](#). We selected two cut-points (yielding three pieces) at times that ensured a roughly equal increment in the cumulative distribution function. With this strategy we anticipate enough transitions to ensure the baseline intensities are reasonably well estimated.

Given cut-points and letting $d_{kh}(t) = I(t \in \mathcal{B}_{kh})$, $h = 1, \dots, q_k$, $k = 1, \dots, K$, we can write

$$\lambda_{k0}(t) = \prod_{h=1}^{q_k} \lambda_{kh}^{d_{kh}(t)},$$

and

$$\Lambda_k(s, t \mid \mathbf{x}) = \sum_{h=1}^{q_k} u_{kh}(s, t) \lambda_{kh} \exp(\mathbf{x}' \boldsymbol{\beta}_k), \text{ for } s < t,$$

where $u_{kh}(s, t) = \int_s^t d_{kh}(u) du = \max(\min(b_{kh}, t) - \max(b_{k,h-1}, s), 0)$. If $\boldsymbol{\lambda}_k = (\lambda_{k1}, \dots, \lambda_{kq_k})'$, $k = 1, \dots, K$, $\boldsymbol{\lambda} = (\boldsymbol{\lambda}'_1, \dots, \boldsymbol{\lambda}'_K)'$, $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_K)'$, let $\boldsymbol{\theta} = (\boldsymbol{\lambda}', \boldsymbol{\beta}')$.

For a sample of n independent processes let t_{ik} be the time of entry to state k for subject i , $k = 1, \dots, K$, and $t_{i0} = 0$, $i = 1, \dots, n$. With complete data $\{t_{ik}, k = 1, \dots, K, \mathbf{x}_i, i =$

$1, \dots, n\}$ the complete data log-likelihood takes the form

$$\log \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{k=1}^K \left[\sum_{h=1}^{q_k} d_{kh}(t_{ik}) \log \lambda_{kh} + \mathbf{x}'_i \boldsymbol{\beta}_k - \sum_{h=1}^{q_k} u_{kh}(t_{i,k-1}, t_{ik}) \lambda_{kh} \exp(\mathbf{x}'_i \boldsymbol{\beta}_k) \right], \quad (2.5)$$

Note that $d_{kh}(t)$ is simply an indicator of whether t falls in the region $\mathcal{B}_{kh}$ so $d_{kh}(t_{ik})$ indicates whether state k was entered during interval $\mathcal{B}_{kh}$ for individual i , $i = 1, \dots, n$. Likewise $u_{kh}(t_{i,k-1}, t_{ik})$ is the time at risk of a transition into state k in interval $\mathcal{B}_{kh}$ for individual i , $h = 1, \dots, q_k, k = 1, \dots, K, i = 1, \dots, n$. With complete data, all the terms $\{(d_{kh}(t_{ik}), u_{kh}(t_{i,k-1}, t_{ik})), h = 1, \dots, q_k, k = 1, \dots, K\}$ are observed for individual i , $i = 1, \dots, n$, and (2.5) can be maximized using software for fitting generalized linear models since it has the form of a log-likelihood for a Poisson process (Cook and Lawless, 2007). When processes are under intermittent observation this complete data log-likelihood is the basis of the EM algorithm (Dempster and others, 1977) where we aim to obtain the conditional expectations of $\{(d_{kh}(T_{ik}), u_{kh}(T_{i,k-1}, T_{ik})), h = 1, \dots, q_k, k = 1, \dots, K\}$ given the observed data $\mathcal{D}_i = \{a_{ij}, Z(a_{ij}), j = 0, 1, \dots, m_i, \mathbf{x}_i\}$. We describe how to do this in the following section.

2.3.2 The Expectation-Maximization Algorithm

To maximize the target function (2.4), we adopt an Expectation-Maximization (EM) algorithm (Dempster and others, 1977). In what follows we describe how the objective function at the maximization step can be reformulated into one that has the same form as the log-likelihood of a time homogeneous Poisson process which means that existing software for variable selection can be employed.

Initialization

The initial value $\boldsymbol{\theta}^{(0)}$ is determined by first setting $\boldsymbol{\beta}_k = \mathbf{0}$, $k = 1, \dots, K$. For the baseline intensities we fit a null multistate model (i.e. with no covariates) with time-homogeneous baseline intensities and use the resulting maximum likelihood estimates as the starting values for these parameters. This model can be fitted in R using the `msm` function by [Jackson \(2011\)](#).

The Expectation Step

The penalized complete data log-likelihood is obtained by subtracting the penalty term $p_\gamma(\boldsymbol{\beta})$ from the complete data log-likelihood (2.5). This is given by

$$\ell_p(\boldsymbol{\theta}) = \frac{1}{n} \log \mathcal{L}(\boldsymbol{\theta}) - p_\gamma(\boldsymbol{\beta}),$$

where $p_\gamma(\boldsymbol{\beta})$ takes one of the forms discussed in Section 2.2.2. The observed data for all subjects is denoted by $\mathcal{D} = \{\mathcal{D}_i, i = 1, \dots, n\}$ so at the r th step with interim estimate $\boldsymbol{\theta}^{(r)}$, we aim to compute

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) = E \{ \ell_p(\boldsymbol{\theta}) \mid \mathcal{D}; \boldsymbol{\theta}^{(r)} \} . \quad (2.6)$$

Letting $\delta_{ikh}^{(r)} = E[d_{kh}(T_{ik}) \mid \mathcal{D}_i; \boldsymbol{\theta}^{(r)}]$ and $\mu_{ikh}^{(r)} = E[u_{kh}(T_{i,k-1}, T_{ik}) \mid \mathcal{D}_i; \boldsymbol{\theta}^{(r)}]$, this can be written as

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \sum_{h=1}^{q_k} \left[\delta_{ikh}^{(r)} (\log \lambda_{kh} + \mathbf{x}'_i \boldsymbol{\beta}_k) - \mu_{ikh}^{(r)} \lambda_{kh} \exp(\mathbf{x}'_i \boldsymbol{\beta}_k) \right] - p_\gamma(\boldsymbol{\beta}) . \quad (2.7)$$

Note that because of the form of (2.7) and the fact that the missing terms in $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)})$ are additive, the conditional expectation of each missing quantity can be computed individually.

Computing Conditional Expectations

The complete data log-likelihood involves statistics

$$\{(d_{kh}(t_{ik}), u_{kh}(t_{i,k-1}, t_{ik}), h = 1, \dots, q_k, k = 1, \dots, K, i = 1, \dots, n)\},$$

for the parameters of the conceptual Poisson-type log-likelihood in (2.5). Here we discuss how to calculate the conditional expectations of these terms given the observed data - this must be carried out in each E-step of the EM algorithm. Of course the nature of the calculations is the same for all individuals, so in what follows we drop the subscript i indexing individuals and discuss the calculations for a single individual for whom we observe, say, $\mathcal{D} = \{(a_j, Z(a_j)), j = 1, 2, \dots, m, \mathbf{x}\}$. Let $\mathcal{C}_k = (L_k, R_k]$ denote the shortest interval known to contain T_k where $L_k = \max\{a_j : Z(a_j) < k\}$ and $R_k = \min\{a_j : Z(a_j) \geq k\}$ if $Z(a_m) \geq k$ and $R_k = \infty$ otherwise. If we let z_{Lk} and z_{Rk} denote the observed state occupied at L_k and R_k respectively, the data necessary to compute the conditional expectations for $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)})$ is simply $\mathcal{D}^k = \{T_k \in \mathcal{C}_k, Z(L_k) = z_{Lk}, Z(R_k) = z_{Rk}, \mathbf{x}'\}$, $k = 1, \dots, K$. If $\boldsymbol{\theta}^{(r)}$ is the estimate of $\boldsymbol{\theta}$ at the r th iteration then the required conditional expectation of $d_{kh}(T_k)$ has the form

$$\delta_{kh}^{(r)} = E(d_{kh}(T_k) \mid \mathcal{D}^k; \boldsymbol{\theta}^{(r)}) = E\{I(T_k \in \mathcal{B}_{kh}) \mid T_k \in \mathcal{C}_k, Z(L_k), Z(R_k), \mathbf{x}; \boldsymbol{\theta}^{(r)}\},$$

and is expressed as

$$\delta_{kh}^{(r)} = \frac{P(Z(L_k) = z_{Lk}, Z(b_{k,h-1}) = k-1, Z(b_{kh}) = k, Z(R_k) = z_{Rk} \mid Z(a_0) = 0, \mathbf{x}; \boldsymbol{\theta}^{(r)})}{P(Z(L_k) = z_{Lk}, Z(R_k) = z_{Rk} \mid Z(a_0) = 0, \mathbf{x}; \boldsymbol{\theta}^{(r)})}. \quad (2.8)$$

The numerator in (2.8) is the probability of all possible paths consistent with the relevant observed data and $T_k \in \mathcal{B}_{kh}$ while the denominator is the probability that $T_k \in \mathcal{C}_k$. The

conditional expectation of the time in state $k - 1$ at risk of a transition into state k during interval $\mathcal{B}_{kh}$ given the observed data, is computed at the r th iteration as

$$\mu_{kh}^{(r)} = \int_{\mathcal{B}_{kh}} E\{I(Z(s) = k - 1) \mid \mathcal{D}; \boldsymbol{\theta}^{(r)}\} ds,$$

which can be written as

$$\mu_{kh}^{(r)} = \int_{\mathcal{B}_{kh}} \frac{P(Z(L_k) = z_{Lk}, Z(b_{k,h-1}) = k - 1, Z(b_{kh}) = k, Z(R_k) = z_{Rk} \mid Z(a_0) = 0, \mathbf{x}; \boldsymbol{\theta}^{(r)})}{P(Z(L_k) = z_{Lk}, Z(R_k) = z_{Rk} \mid Z(a_0) = 0, \mathbf{x}; \boldsymbol{\theta}^{(r)})} ds.$$

The Maximization Step

Here we show how the M-step can be reformulated to align with the optimization problem of a penalized Poisson likelihood so that we can exploit existing software. First we let $v_{khr} = I(h = r)$ for $h, r = 1, \dots, q_k$ and define the $q_k \times 1$ vector $\mathbf{v}_{kh} = (v_{kh1}, \dots, v_{khq_k})'$. Then let $w_{kl} = I(k = l)$, $k, l = 1, \dots, K$ and construct the pseudo-covariate $\bar{\mathbf{x}}_{ikh} = (w_{k1}\mathbf{v}'_{1h}, w_{k2}\mathbf{v}'_{2h}, \dots, w_{kK}\mathbf{v}'_{Kh}, w_{k1}\mathbf{x}'_i, w_{k2}\mathbf{x}'_i, \dots, w_{kK}\mathbf{x}'_i)'$. If $\alpha_{kh} = \log \lambda_{kh}$, $\boldsymbol{\alpha}_k = (\alpha_{k1}, \dots, \alpha_{kq_k})'$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}'_1, \dots, \boldsymbol{\alpha}'_K)'$, then let $\boldsymbol{\theta} = (\boldsymbol{\alpha}'_1, \dots, \boldsymbol{\alpha}'_K, \boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_K)'$. We can then rewrite (2.7) as

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \sum_{h=1}^{q_k} \left[\delta_{ikh}^{(r)} \bar{\mathbf{x}}'_{ikh} \boldsymbol{\theta} - \mu_{ikh}^{(r)} \exp(\bar{\mathbf{x}}'_{ikh} \boldsymbol{\theta}) \right] - p_\gamma(\boldsymbol{\beta}). \quad (2.9)$$

This takes the form of a Poisson log-likelihood with pseudo-individuals labeled by (i, k, h) and penalized with respect to a subset of the parameters $\boldsymbol{\beta}$ within $\boldsymbol{\theta}$. Specifically the response for pseudo-individual (i, k, h) is $\delta_{ikh}^{(r)}$, the covariate vector is $\bar{\mathbf{x}}_{ikh}$, and $\mu_{ikh}^{(r)}$ as the duration of time at risk, leading to an offset $\log \mu_{ikh}^{(r)}$. This coding of the pseudo-

covariate means that covariates $w_{k1}\mathbf{v}_{1h}, \dots, w_{kK}\mathbf{v}_{Kh}$ are controlled for to ensure the baseline intensities remain flexible (weakly parametric) but selection will be carried out for the remaining covariates $w_{k1}\mathbf{x}_i, \dots, w_{kK}\mathbf{x}_i$ due to the presence of the penalty applied to β . The benefit of this reformulation is that once a dataset is created in this manner one can use existing R functions for variable selection based on penalized likelihood for Poisson regression.

Maximization of the penalized Poisson likelihood with LASSO or adaptive LASSO can be implemented using the `glmnet` package in R by specifying the family argument as `poisson`. For adaptive LASSO, adaptive weights can be set using the `penalty.factor` argument. The `grplasso` package in R can be used to perform penalized Poisson regression with group LASSO or adaptive group LASSO. However, `grplasso` does not support adaptive weights for coefficients. We modified source code to introduce a weight argument and enable the use of adaptive weights for groups. Details on this modification and R code for the implementation of the EM algorithm is provided on <https://github.com/XianweiLi33/MS-penalized>.

2.4 Simulation Studies

In this section, we conduct simulation studies to assess the performance of the proposed penalized EM algorithm for variable selection with multistate data under intermittent observation and compare the performance of several penalty functions.

2.4.1 Design of Simulation Studies

Here we consider a three state progressive process with $K = 2$ and assume the covariate vector $\mathbf{X}$ follows a p -dimensional multivariate normal distribution $MVN_p(\mathbf{0}, \Sigma)$ with $p = 100$ and Σ having an autoregressive dependence structure with (j, k) entry $\Sigma_{jk} = \kappa^{|j-k|}$, $j, k = 1, \dots, 100$, and $\kappa = 0.5$. We investigate three different settings each defined by how the covariates influence the transition intensities, as described below. In Setting 1 covariates $x_1, \dots, x_5$ and $x_{95}, \dots, x_{100}$ have strong positive effects on both the $0 \rightarrow 1$ and $1 \rightarrow 2$ transition intensities while the remaining 90 variables $x_6, \dots, x_{94}$ have no effects (i.e. are “noise” variables). Here we can examine how well the covariates are selected overall and for the respective transition intensities. In Setting 2 covariates $x_1, \dots, x_5$ affect the $1 \rightarrow 2$ transition but not the $0 \rightarrow 1$ transition. We consider this to gauge how the algorithm performs when covariates may only affect part of the disease process. The strengths of effects of important covariate effects may differ across transitions so in Setting 3, covariates $x_1, \dots, x_5$ and $x_{95}, \dots, x_{100}$ have strong effects on the $0 \rightarrow 1$ transition, but weak effects on the $1 \rightarrow 2$ transition (again covariates $x_6, \dots, x_{94}$ have no effects). The particular parameter values are given in the footnotes of the tables reporting the simulation results.

The follow-up period was $(0, \tau]$ which ends with administrative censoring time $\tau = 1$. The baseline intensities for the $0 \rightarrow 1$ and $1 \rightarrow 2$ transition are time homogeneous and denoted by λ_{10} and λ_{20} , respectively. We further set $\eta = \lambda_{20}/\lambda_{10} = 1.5$ and solve for λ_{10} and λ_{20} such that $E[P(T_2 < \tau \mid \mathbf{X}; \boldsymbol{\theta})] = 0.95$. The event state entry times T_1 and T_2 are generated given the intensity model $\lambda_k(t \mid \mathbf{x}) = \lambda_{k0} \exp(\mathbf{x}'\boldsymbol{\beta}_k)$. We adopt the piecewise constant baseline intensities as described earlier assuming a common set of cut-points for the two transitions; cut-points are set to 0.2 and 0.4. Visits times are generated according

to a time-homogeneous Poisson process with rate ρ ; we set the average number of visits over $(0, 1]$ to $\rho = 16$ or 4 . To generate the visit times we first generate the number of visits for an individual as $M \sim \text{Poisson}(\rho)$ and then given $M = m$ the visit times $a_1, a_2, \dots, a_m$ are generated as uniform random variables over $[0, 1]$. After recording the states occupied at each visit time for each individual, we have the multistate data under intermittent observation. For each setting we generate samples of size $n = 500$ or 1000 and using i to index individuals we write the data as $\{(a_{ir}, Z_i(a_{ir})), r = 0, \dots, m_i, \mathbf{x}_i; i = 1, \dots, n\}$ where $a_{i0} = 0$ and $Z_i(a_{i0}) = 0$.

We compare the performance of the EM algorithm with various penalty functions including the LASSO, adaptive LASSO (ALASSO), group LASSO (GLASSO), and adaptive group LASSO (AGLASSO) in the simulation studies. With the adaptive LASSO penalty, the penalty term takes the form

$$\gamma \sum_{k=1}^2 \sum_{j=1}^{100} w_{kj}^{(r)} |\beta_{kj}|,$$

where $w_{kj}^{(r)} = 1/|\tilde{\beta}_{kj}^{(r-1)}|$ and $\tilde{\beta}_{kj}^{(r-1)}$ is the estimator in the $(r-1)$ th iteration. The adaptive weights are updated in each iteration as done in [Wu and Cook \(2015\)](#). We take $\tilde{\beta}_{kj}^{(0)}$ as the MLE which maximizes the unpenalized observed log-likelihood (2.3). $\tilde{\beta}_{kj}^{(0)}$ can be obtained by running the proposed EM algorithm without a penalty term. Similarly, with the group LASSO penalty, the penalty term takes the form

$$\gamma \sum_{j=1}^{100} w_j^{(r)} \sqrt{\beta_{1j}^2 + \beta_{2j}^2},$$

where $w_j^{(r)} = 1/\sqrt{(\tilde{\beta}_{1j}^{(r-1)})^2 + (\tilde{\beta}_{2j}^{(r-1)})^2}$ and $\tilde{\beta}_{kj}^{(r-1)}$ is the estimator in the $(r-1)$ th iteration.

Cross-validation (CV) and generalized cross-validation (GCV) are popular approaches for choosing the regularization parameters. [Wang and others \(2007\)](#) pointed out that with the smooth clipped absolute deviation (SCAD) penalty, models selected by GCV or AIC tend to over-fit and [Zhang and others \(2010\)](#) showed that BIC-type selectors can select the true model consistently with the SCAD penalty. [Wang and others \(2007\)](#) also applied BIC-type selectors with adaptive LASSO penalty and proved its selection consistency property; [Li and others \(2020\)](#) confirmed selection consistency for the adaptive group LASSO with this selection algorithm. It is important to note that the form of BIC is an approximation of the likelihood which neglects a residual term and therefore can be inaccurate when the sample size is moderate or small. Many variants of the BIC have been proposed to address this problem; see [Sclove \(1987\)](#) and [Chen and Chen \(2008\)](#). [Chaurasia \(2023\)](#) defined a class of BIC variants (a so-called gBIC) and showed that the Sample Adjusted BIC (SABIC) ([Sclove, 1987](#)) is a special case of the gBIC and equivalent to BIC when the sample size goes to infinity. We select the tuning parameter γ using the SABIC which has the form

$$\text{SABIC}(\gamma) = -2\ell(\hat{\boldsymbol{\theta}}_\gamma) + |S_\gamma| \log \frac{n+2}{24},$$

where $\hat{\boldsymbol{\theta}}_\gamma$ is the estimate obtained by maximizing the penalized observed likelihood (2.4) given γ , $\ell(\hat{\boldsymbol{\theta}}_\gamma)$ is the unpenalized observed log-likelihood (2.3) evaluated at $\hat{\boldsymbol{\theta}}_\gamma$, S_γ is the set of non-zero estimates of the coefficients, $|S_\gamma|$ is the size of the set, and n is the sample size. The convergence criterion for the EM algorithm is $\max_{1 \leq j \leq 206} |(\boldsymbol{\theta}_j^{(r+1)} - \boldsymbol{\theta}_j^{(r)})/\boldsymbol{\theta}_j^{(r)}| < 10^{-6}$, where $\boldsymbol{\theta}_j^{(r)}$ is the j th element of $\boldsymbol{\theta}$ at the r th iteration and $\boldsymbol{\theta}$ contains 206 parameters in total; there are 100 covariates with two potential coefficients each, and three pieces for each of the two baseline intensities. We set the maximum number of iterations for the EM algorithm to 100 and over 98% of the implementations of the algorithm converged for the LASSO and ALASSO penalty functions - this number was somewhat lower for the penalty

functions using grouping but in all cases a minimum value of the tuning parameter could be identified. We also consider the performance of simple methods for failure time analysis that one might consider in this setting. If interest lies in identification of risk factors for entry to the absorbing state 2 (e.g. onset of axial disease in our motivating study) one could construct a dataset for the interval-censored state 2 entry time T_2 . Oftentimes the right endpoint of the censoring interval is taken as the failure time as it may correspond to the time of diagnosis of a condition; for example, in the motivating problem this would correspond to the first time the patient was known to have axial disease. The mid-point is also often adopted as an ad hoc estimate of the time of the event of interest ([Lindsey and Ryan, 1998](#)). We refer to these as right-endpoint and mid-point imputation strategies, respectively. Following imputation, a Cox regression model can be fitted with the LASSO or ALASSO penalty functions; as there is only one set of regression coefficients in this case, there is no role for the “grouped” versions of these penalty functions.

2.4.2 Performance Metrics

We report the results over $n_{sim} = 100$ independent samples in what follows. The performance of variable selection is assessed by true positive and false positive rates defined in ways that reflect the uniqueness of the multistate setting. For the covariates $x_1, \dots, x_{100}$, there are two coefficients related to each covariate, representing the effects on the $0 \rightarrow 1$ and $1 \rightarrow 2$ transitions, respectively. Since a given covariate may exert an effect of no, one or two intensities we use the term selection of effects rather than selection of variables. That is a variable may be selected and found to have an effect on one transition but not the other; sometimes it may be of interest to record how often it was selected for the correct transition intensity and sometimes it may be of interest to simply record whether it was

selected at all if it has some role in the disease process.

True Positive Rates

Let $N_k = \sum_{j=1}^p I(\beta_{kj} \neq 0)$ be the number of covariates with an effect on the $k - 1 \rightarrow k$ transition. This can be equivalently described as the number of truly non-zero regression coefficients in the $k - 1 \rightarrow k$ transition intensity. Then let $N = \sum_{k=1}^2 N_k$ denote the number of non-zero coefficients across $0 \rightarrow 1$ and $1 \rightarrow 2$ transitions and covariates. Then let $N_k^D = \sum_{j=1}^p I(\beta_{kj} \neq 0)I(\hat{\beta}_{kj} \neq 0)$ be the number of truly non-zero coefficients which were detected as non-zero for the $k - 1 \rightarrow k$ transition and $N^D = \sum_{k=1}^2 N_k^D$ denote the total number of truly non-zero coefficients which were estimated to be non-zero (and hence correspond to effects that are correctly detected). Based on these summary statistics the transition-specific true positive rate is $FP_k = N_k^D/N_k$ and the overall true positive rate is $FP = N^D/N$. One may also be interested in assessing the probability of correctly identifying covariates as important among those having any role in the disease process; this is not concerned with whether the variable was identified as affecting the correct intensity but just whether a variable having some effect in the disease process is found to have any effect. Let $N_{\text{cov}} = \sum_{j=1}^p I(\beta_{1j} \neq 0 \text{ or } \beta_{2j} \neq 0)$ denote the number of covariates that have some effect on at least one transition intensity, and $N_{\text{cov}}^D = \sum_{j=1}^p I(\beta_{1j} \neq 0 \text{ or } \beta_{2j} \neq 0)I(\hat{\beta}_{1j} \neq 0 \text{ or } \hat{\beta}_{2j} \neq 0)$ denote the number of such variables that are found to have one or more effects. Then $TP_{\text{cov}} = N_{\text{cov}}^D/N_{\text{cov}}$ represents the proportion of important (in some way) covariates that are selected. For the Cox regression analyses following right or mid-point imputation we let $N_k^{D_{\text{cox}}} = \sum_{j=1}^p I(\beta_{kj} \neq 0)I(\hat{\beta}_j \neq 0)$ be the number of truly non-zero effects which were detected to be non-zero for the $k - 1 \rightarrow k$ transition, where $\hat{\beta}_j$ denotes the estimated effect for the j th covariate obtained from the Cox model. The transition-specific true positive rate is given by $TP_k^{\text{cox}} = N_k^{D_{\text{cox}}}/N_k$. Let $N_{\text{cov}}^{D_{\text{cox}}} = \sum_{j=1}^p I(\beta_{1j} \neq 0 \text{ or } \beta_{2j} \neq 0)I(\hat{\beta}_j \neq 0)$

be the number of covariates that have some effects on at least on transition intensity and correctly identified by the Cox model. Then $TP_{cov}^{cox} = N_{cov}^{D_{cox}}/N_{cov}$ represents the proportion of correctly selected covariates.

False Positive Rates

Let $\bar{N}_k = \sum_{j=1}^p I(\beta_{kj} = 0)$ be the number of covariates with no effect for the $k - 1 \rightarrow k$ transition and $\bar{N}_k^D = \sum_{j=1}^p I(\beta_{kj} = 0)I(\hat{\beta}_{kj} \neq 0)$ be the number of covariates with no effect which were incorrectly found to have an effect. The transition-specific false positive rate assesses the probability of falsely selecting variables for the transition from $k - 1$ to k and is given by $FP_k = \bar{N}_k^D/\bar{N}_k$. The overall false positive rate reflects the probability of falsely identifying effects for which there are none and has the form $FP = \bar{N}^D/\bar{N}$ where $\bar{N}^D = \sum_{k=1}^2 \bar{N}_k^D$ and $\bar{N} = \sum_{k=1}^2 \bar{N}_k$. One may be interested in the risk of identifying variables with no effects as important variables. To this end we let $\bar{N}_{cov} = \sum_{j=1}^p I(\beta_{1j} = 0 \text{ and } \beta_{2j} = 0)$ denote the number of variables with no effect whatsoever. and $\bar{N}_{cov}^D = \sum_{j=1}^p I(\beta_{1j} = 0 \text{ and } \beta_{2j} = 0)I(\hat{\beta}_{1j} \neq 0 \text{ or } \hat{\beta}_{2j} \neq 0)$ denote the number of such variables which were incorrectly found to have some effect. This covariate false positive rate is defined as $FP_{cov} = \bar{N}_{cov}^D/\bar{N}_{cov}$. The averaged values of the FP, FP_k , FP_{cov} , and FP_{cov}^{cox} over simulations are reported. For Cox regression analyses we let $\bar{N}_k^{D_{cox}} = \sum_{j=1}^p I(\beta_{kj} = 0)I(\hat{\beta}_j \neq 0)$ be the number of zero effects which were falsely detected to be non-zero for the $k - 1 \rightarrow k$ transition by the Cox model. Then $FP_k^{cox} = \bar{N}_k^{D_{cox}}/\bar{N}_k$ represents the transition-specific false positive rate. Let $\bar{N}_{cov}^{D_{cox}} = \sum_{j=1}^p I(\beta_{1j} = 0 \text{ and } \beta_{2j} = 0)I(\hat{\beta}_j \neq 0)$ be the number of covariates that have no effects on any transition intensities but are falsely identified as having non-zero effects by the Cox model. Then $FP_{cov}^{cox} = \bar{N}_{cov}^{D_{cox}}/\bar{N}_{cov}$ represents the proportion of correctly selected covariates.

2.4.3 Empirical Results

The average of the true positive and false positive rates are reported across 100 simulations in Table 2.1 for a time-homogeneous visit intensity $\rho = 16$ in samples of size $n = 1000$ individuals for each of Settings 1 to 3 described in Section 2.4.1. Table 2.2 gives the corresponding results with less frequent visits where $\rho = 4$ and corresponding results for samples of size $n = 500$ are given in Tables A.1 and A.2 of Appendix A.1. The initial interpretation is based on findings regarding the identification (correct or not) of covariates that have a role in the disease process, which means we emphasize the TP_{cov} and FP_{cov} rates along with the corresponding Cox model rates.

We first remark on the findings regarding the proposed algorithm across different settings and penalty functions; see results under “Multistate Analysis - Penalized EM” in Tables 2.1 and 2.2. For Table 2.1 where $n = 1000$ and visits are quite frequent ($\rho = 16$) the ALASSO and AGLASSO penalty functions lead to better performance than their non-adaptive counterparts in that they have lower false positive rates with an only modest reduction in the true positive rate across all three settings. There is a degradation in the overall true positive rate (TP) and TP_2 for the LASSO and ALASSO in Setting 3 where covariate effects are small for the $1 \rightarrow 2$ transition intensity. The group-based penalties retain good performance in this setting with a high true positive rate and low false positive rate, showing the appeal of the GLASSO and AGLASSO penalty functions in the multistate setting. In terms of the penalized Cox regression based on right- and mid-point imputation, the true positive rates are often quite high and the false positive rates quite low in Table 2.1 but there can be a slight elevation in the false positive rate in Setting 3. In Table 2.2 visits are less frequent ($\rho = 4$), and we see the benefit of the group-based penalty functions – they yield much better true positive rates and lower false positive rates

Multistate Analysis - Penalized EM Algorithm					Penalized Cox Regression Following Imputation				
Metrics	LASSO	ALASSO	GLASSO	AGLASSO	Metrics	Right-endpoint		Mid-point	
						LASSO	ALASSO	LASSO	ALASSO
<i>Setting 1</i> *									
TP	99.9	98.4	100.0	100.0	—	—	—	—	—
TP ₁	100.0	99.9	100.0	100.0	TP ₁ ^{cox}	100.0	99.9	100.0	100.0
TP ₂	99.7	96.9	100.0	100.0	TP ₂ ^{cox}	100.0	99.9	100.0	100.0
TP _{cov}	100.0	100.0	100.0	100.0	TP _{cov} ^{cox}	100.0	99.9	100.0	100.0
FP	6.9	3.3	2.9	1.4	—	—	—	—	—
FP ₁	7.5	2.4	2.9	1.4	FP ₁ ^{cox}	6.7	4.8	6.6	4.4
FP ₂	6.4	4.3	2.9	1.4	FP ₂ ^{cox}	6.7	4.8	6.6	4.4
FP _{cov}	13.3	6.5	2.9	1.4	FP _{cov} ^{cox}	6.7	4.8	6.6	4.4
<i>Setting 2</i> †									
TP	99.5	99.1	99.8	98.9	—	—	—	—	—
TP ₁	100.0	100.0	100.0	100.0	TP ₁ ^{cox}	100.0	100.0	100.0	100.0
TP ₂	99.2	98.6	99.7	98.4	TP ₂ ^{cox}	95.9	85.8	96.4	86.2
TP _{cov}	99.8	99.6	99.7	98.4	TP _{cov} ^{cox}	95.9	85.8	96.4	86.2
FP	6.4	3.4	5.3	4.2	—	—	—	—	—
FP ₁	7.4	3.1	7.8	6.6	FP ₁ ^{cox}	9.8	8.6	9.9	8.6
FP ₂	5.3	3.7	2.7	1.6	FP ₂ ^{cox}	5.3	5.1	5.3	5.1
FP _{cov}	12.1	6.6	2.7	1.6	FP _{cov} ^{cox}	5.3	5.1	5.3	5.1
<i>Setting 3</i> ‡									
TP	70.2	71.8	100.0	99.4	—	—	—	—	—
TP ₁	100.0	100.0	100.0	99.4	TP ₁ ^{cox}	99.9	97.0	100.0	97.8
TP ₂	40.3	43.5	100.0	99.4	TP ₂ ^{cox}	99.9	97.0	100.0	97.8
TP _{cov}	100.0	100.0	100.0	99.4	TP _{cov} ^{cox}	99.9	97.0	100.0	97.8
FP	3.9	3.4	2.7	2.0	—	—	—	—	—
FP ₁	3.8	3.0	2.7	2.0	FP ₁ ^{cox}	5.7	5.7	5.7	5.0
FP ₂	4.0	3.8	2.7	2.0	FP ₂ ^{cox}	5.7	5.7	5.7	5.0
FP _{cov}	7.6	6.7	2.7	2.0	FP _{cov} ^{cox}	5.7	5.7	5.7	5.0

* Setting 1: $x_1, \dots, x_5$ and $x_{96}, \dots, x_{100}$ have strong effects of $\log(1.2)$ on both transitions, while other covariates are noise variables;

† Setting 2: $x_1, \dots, x_5$ have strong effects of $\log(1.2)$ on the $0 \rightarrow 1$ transition but no effects on the $1 \rightarrow 2$ transition; $x_{96}, \dots, x_{100}$ have strong effects of $\log(1.2)$ on both transitions, while other covariates are noise variables;

‡ Setting 3: $x_1, \dots, x_5$ and $x_{96}, \dots, x_{100}$ have strong effects of $\log(1.2)$ on $0 \rightarrow 1$ transition and weak effects of $\log(1.05)$ on $1 \rightarrow 2$ transition, while other covariates are noise variables.

Table 2.1: Empirical results for multistate state analysis under independent intermittent observation scheme summarizing transition-specific and covariate-wise true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression following ad hoc right-endpoint or mid-point imputation; $n = 1000$ with $\rho = 16$ as the mean number of visits over $(0, 1]$.

					Penalized Cox Regression Following Imputation				
Multistate Analysis - Penalized EM Algorithm					Right-endpoint		Mid-point		
Metrics	LASSO	ALASSO	GLASSO	AGLASSO	Metrics	LASSO	ALASSO	LASSO	ALASSO
<i>Setting 1</i> *									
TP	94.4	86.5	100.0	99.0	—	—	—	—	—
TP ₁	100.0	99.3	100.0	99.0	TP ₁ ^{cox}	96.8	83.8	100.0	95.3
TP ₂	88.7	73.7	100.0	99.0	TP ₂ ^{cox}	96.8	83.8	100.0	95.3
TP _{cov}	100.0	99.9	100.0	99.0	TP _{cov} ^{cox}	96.8	83.8	100.0	95.3
FP	5.4	2.4	2.5	0.6	—	—	—	—	—
FP ₁	7.0	1.3	2.5	0.6	FP ₁ ^{cox}	4.6	6.2	5.1	5.5
FP ₂	3.7	3.5	2.5	0.6	FP ₂ ^{cox}	4.6	6.2	5.1	5.5
FP _{cov}	10.4	4.8	2.5	0.6	FP _{cov} ^{cox}	4.6	6.2	5.1	5.5
<i>Setting 2</i> †									
TP	94.3	90.3	99.8	92.6	—	—	—	—	—
TP ₁	99.8	99.4	100.0	99.6	TP ₁ ^{cox}	99.0	91.8	100.0	98.6
TP ₂	91.5	85.8	99.7	89.1	TP ₂ ^{cox}	83.0	69.4	91.2	78.7
TP _{cov}	97.2	94.3	99.7	89.1	TP _{cov} ^{cox}	83.0	69.4	91.2	78.7
FP	4.6	2.4	5.3	3.3	—	—	—	—	—
FP ₁	5.9	2.2	7.8	5.2	FP ₁ ^{cox}	7.1	9.2	8.3	7.9
FP ₂	3.2	2.6	2.7	1.2	FP ₂ ^{cox}	3.8	7.1	4.2	5.1
FP _{cov}	8.8	4.5	2.7	1.2	FP _{cov} ^{cox}	3.8	7.1	4.2	5.1
<i>Setting 3</i> ‡									
TP	57.9	58.6	99.9	95.7	—	—	—	—	—
TP ₁	99.8	97.9	99.9	95.7	TP ₁ ^{cox}	92.1	76.9	99.2	89.3
TP ₂	15.9	19.2	99.9	95.7	TP ₂ ^{cox}	92.1	76.9	99.2	89.3
TP _{cov}	99.9	98.2	99.9	95.7	TP _{cov} ^{cox}	92.1	76.9	99.2	89.3
FP	2.8	2.4	2.2	1.1	—	—	—	—	—
FP ₁	3.4	2.4	2.2	1.1	FP ₁ ^{cox}	4.1	6.2	4.9	5.9
FP ₂	2.2	2.4	2.2	1.1	FP ₂ ^{cox}	4.1	6.2	4.9	5.9
FP _{cov}	5.5	4.7	2.2	1.1	FP _{cov} ^{cox}	4.1	6.2	4.9	5.9

* Setting 1: $x_1, \dots, x_5$ and $x_{96}, \dots, x_{100}$ have strong effects of $\log(1.2)$ on both transitions, while other covariates are noise variables;

† Setting 2: $x_1, \dots, x_5$ have strong effects of $\log(1.2)$ on the $0 \rightarrow 1$ transition but no effects on the $1 \rightarrow 2$ transition; $x_{96}, \dots, x_{100}$ have strong effects of $\log(1.2)$ on both transitions, while other covariates are noise variables;

‡ Setting 3: $x_1, \dots, x_5$ and $x_{96}, \dots, x_{100}$ have strong effects of $\log(1.2)$ on $0 \rightarrow 1$ transition and weak effects of $\log(1.05)$ on $1 \rightarrow 2$ transition, while other covariates are noise variables.

Table 2.2: Empirical results for multistate state analysis under independent intermittent observation scheme summarizing transition-specific and covariate-wise true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression following ad hoc right-endpoint or mid-point imputation; $n = 1000$ with $\rho = 4$ as the mean number of visits over $(0, 1]$.

than the corresponding non-grouped counterparts. Moreover we see generally lower false positive rates from the proposed penalized EM algorithm compared to penalized Cox regression here. When the sample size is $n = 500$ (see Table A.1) the differences between the performance of the multistate and Cox regression analyses are more pronounced with the latter typically (but not always) giving lower true positive rates and higher false positive rates.

A unique aspect of variable selection with multistate processes is that covariates may affect some, but not other, intensities – it is with this in mind that we defined transition-specific true positive and false positive rates and we now comment on associated findings. When non-grouped penalties are used the true positive rates for covariates affecting the $0 \rightarrow 1$ transition (TP_1) are consistently higher than those for the covariates affecting the $1 \rightarrow 2$ transition (TP_2) in all settings, with the differences more pronounced when visits are less frequent; see Table 2.2 where $\rho = 4$. For progressive models of this type, fewer $1 \rightarrow 2$ transitions will be known to have occurred over the course of follow-up compared to $0 \rightarrow 1$ transitions, and transitions convey information about marker effects. When non-grouped penalties are used in Setting 3 where the covariate effects are weak for the $1 \rightarrow 2$ transition intensity the TP_2 rate is appreciably lower than it is in Settings 1 and 2 and again the degradation in performance is even more apparent when visits are less frequent. The superior performance at selection (in terms of TP_2) from use of grouped penalties compared to their non-grouped counterparts can be attributed to the fact that they leverage information regarding effects on the $0 \rightarrow 1$ transitions where the times are effectively subject to a milder censoring rate and/or covariate effects may be stronger. The grouping structure can lead to higher false positive rates, however, when covariates have strong effects for one transition and no effect for another. In Setting 2 the first five covariates have no effect on the $0 \rightarrow 1$ transitions but they do affect the $1 \rightarrow 2$ transition.

Here the group LASSO and adaptive group LASSO tend to select more covariates with no effect on the $0 \rightarrow 1$ transitions, resulting in higher FP_1 than those of standard LASSO and adaptive LASSO.

Additional simulations exploring the impact of state-dependent visit intensities (which violate the CIVP assumption (Cook and Lawless, 2021)) are reported on in Appendix A.1 for $n = 1000$ and $\rho = 4$. There we remark that a modest dependence between the disease and visit processes can lead to a slight degradation of the performance (lower true positive and higher false positive rates) but this is relatively minor in large samples.

2.5 HLA Markers and Sacroiliac Joint Damage Data

Here we report the results of applying the proposed penalized EM algorithm to identify HLA markers for the progression of sacroiliac joint damage in PsA patients. The dataset is comprised of 1006 individuals with an average of 4.07 follow-up visits (the 90th percentile is 9).

To examine the extent to which individuals in the sample are seen to enter states 1 and 2, we provide the nonparametric estimate for the cumulative distribution functions for the onset of unilateral damage (entry to state 1) and axial disease (entry to state 2) in the left and right panels of Figure 2.3 respectively. The respective censoring intervals are $\mathcal{C}_1$ and $\mathcal{C}_2$ as discussed in Section 2.1.2. The solid lines are the Turnbull estimates (Turnbull, 1976) obtained from the EMICM package in *R* (Anderson-Bergman, 2017) and the grey lines are pointwise 95% confidence intervals based on 200 nonparametric bootstrap samples. It is apparent that unilateral SI joint damage is quite common after 20 years while axial disease is less common. Based on the estimates of the cumulative distribution functions for the state 1 and 2 entry times, we place the cut-points for the baseline intensities at 10 and

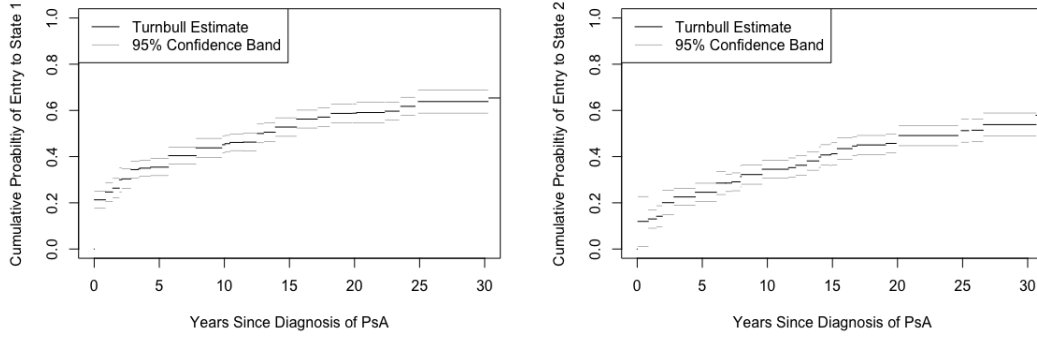


Figure 2.3: Turnbull (Turnbull, 1976) estimates with pointwise 95% confidence intervals the cumulative distribution function of the time to entry to state 1 (left panel) and state 2 (right panel)

20 years; these define intervals with roughly comparable increments in the nonparametric estimates. In the models used for variable selection we control for sex, coded as 1 for males and 0 for females. We select HLA markers associated with the development of SI joint damage from 41 candidate HLA markers with each coded as 1 if present or 0 otherwise. The simulation studies showed that penalties with adaptive weights are sensitive to model misspecification, and penalties with a grouping structure can result in a high false positive rate when the group contains both zero and non-zero effects. We therefore used the LASSO penalty function for this illustration.

Figure 2.4 (a) shows the change in the SABIC over a grid of 11 values of $\log \gamma$ with the optimal value of γ chosen to minimize the SABIC. The algorithm took 21 minutes to converge for each value of γ – these EM algorithms were implemented in parallel to increase efficiency. The right panel in Figure 2.4 shows how estimates shrink as the penalty increases, where the solid lines represent estimates for the $0 \rightarrow 1$ transitions and the dashed lines represent estimates for the $1 \rightarrow 2$ transitions.

Table 2.3 shows a list of the selected variables along with their effects on the transitions. We observe that male sex is associated with an increase in the $0 \rightarrow 1$ transition intensity, suggesting that male patients tend to develop unilateral damage more rapidly than females. While modeling unilateral back involvement has not been carried out in this way in the literature to our knowledge, male sex has been reported as a risk factor for sacroiliac damage (Eder and others, 2013). Indeed since unilateral sacroiliac joint damage is required in order to satisfy bilateral involvement, this result is consistent with these findings. Interestingly while it seems to accelerate the start of the process, there is no evidence that it is associated with progression from unilateral involvement to axial disease. Among the selected HLA markers for the onset of unilateral sacroiliac damage (HLA-B27, HLA-B46, and HLA-C12) all have been reported by others to be associated with sacroiliitis; see Ikumi and others (2019) for example. For progression from unilateral to axial disease, HLA-B41, HLA-C5 and HLA-C17 were selected as having protective effects as the coefficients were all negative. These markers have not been reported to have a role in sacroiliac joint damage but we remark that their interpretation is closely linked to the model we use — a 3-state progressive model which has not been used in this context to our knowledge. In particular, we find that among patients who have developed unilateral sacroiliac joint damage, those who are HLA-B41, HLA-C5, or HLA-C17 positive have a slightly lower intensity for going on to develop axial disease. These new markers, along with HLA-B46 warrant further investigation in external validation studies. The analyses based on right- and mid-endpoint imputation are also provided and the findings here depend on which imputation strategy is adopted. Under right-endpoint imputation HLA-B15, HLA-B27, HLA-B46, HLA-C2 and HLA-C12 are selected as risk factors for axial disease, with HLA-B41 found to be protective. For mid-point imputation HLA-B27, HLA-B46 and HLA-C12 are selected as risk factors for axial disease, in addition to HLA-B14 which is found to be

	Multistate Analysis		Cox regression	
	$0 \rightarrow 1$	$1 \rightarrow 2$	Right-endpoint	Mid-point
Sex	0.191		0.356	0.287
B14				-0.076
B15			0.185	
B27	0.138		0.229	0.262
B41		-0.183	-0.082	
B46	0.367		0.557	0.175
C2			0.016	
C5		-0.001		
C12	0.082		0.147	0.120
C17		-0.204		

Table 2.3: Selected HLA markers for the onset and progression of SI joint damage including unilateral SI joint damage (state 1) and axial disease (state 2) and their effects by multistate penalized EM algorithm and penalized Cox regression after right-endpoint and mid-point imputation using a LASSO penalty

protective given the status of these three markers. We emphasize that we do not intend for readers to interpret the regression coefficients formally from the penalized algorithms; rather, we present the estimates to reflect their relative magnitudes and to quantify the fitted effects at the point where the optimal tuning parameter is selected. In general there is some overlap in the variables selected from the multistate and Cox regression analyses. The multistate analysis has the added benefit of being based on a more nuanced analysis which provides insight into the role the different markers can play in different stages of the disease process. Again the findings from these analyses warrant further exploration in validation studies.

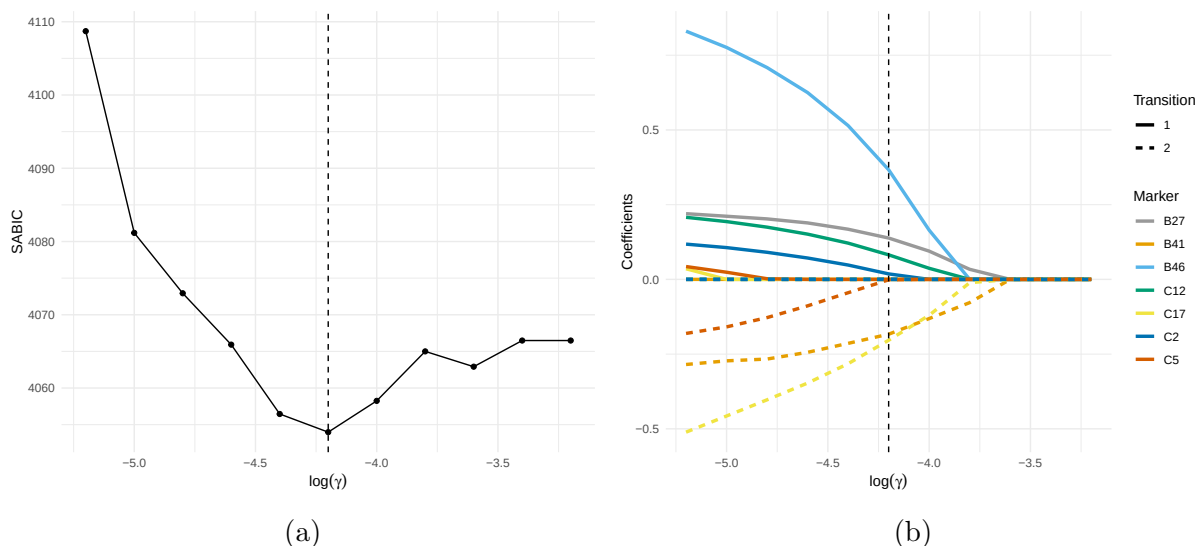


Figure 2.4: (a) Changes in the sample adjusted BIC (SABIC) and (b) shrinkage estimates of coefficients from penalized regression with the LASSO penalty across values of $\log \gamma$

2.6 Discussion

The purpose of this chapter was to develop methods for variable selection for intensity-based models for multistate processes via penalized likelihood. Covariates may affect multiple transition intensities and so in addition to standard penalty functions, penalty functions incorporating grouping of variables can be useful. Here, the covariates are not necessarily grouped according to their role in a disease processes (as may be the case in groups of genes sharing function in protein expression), but rather the coefficients of a given covariate across multiple transitions intensities can be viewed as grouped. The algorithm deals with the setting in which a continuous-time conditional Markov process is observed intermittently. The simplicity of the algorithm arises from the fact that under a piecewise-constant baseline intensity specification the complete data log-likelihood aligns with that of a sum of penalized Poisson likelihoods — this allows exploitation of the `glmnet` function. If the

disease process is subject only to right-censoring, then this relation be exploited directly without the need for an expectation-maximization algorithm.

The computing time for the EM algorithm depends on i) features of the data such as the sample size, the frequency of visits, and the extent of censoring both in terms of the percentage of processes that are not known to enter the absorbing state, and in terms of the frequency of follow-up assessments, ii) the nature of the model such as the number of cut-points used for the baseline intensities and the number of covariates under consideration, and iii) the fineness of the grid used for the selection of the tuning parameter, the convergence criteria, and the choice of penalty function (weighted penalty functions, for example, require fitting an additional model to obtain the weights). For the simulation studies and analyses reported on here we used supercomputing clusters maintained by the Digital Research Alliance of Canada. On this platform in the simulation studies where $n = 1000$ involving an average of 16 visits and a grid of 7 points, the optimal tuning parameter was obtained in approximately 22 minutes for the LASSO penalty function. When the sample size is halved ($n = 500$) or doubled ($n = 2000$), the mean times to convergence were 27 and 40 minutes, respectively. When applying penalties with a group structure, convergence takes slightly longer compared to their non-group counterparts. In addition, the use of ALASSO and AGLASSO requires more time than LASSO and GLASSO due to the need to compute the MLE as the initial estimator for adaptive weights in the first step.

We assumed that the visit process satisfies the CIVP assumption ([Cook and Lawless, 2021](#)) but sensitivity analyses show that there is relatively little effect on the performance if there is a mild violation of the CIVP assumption. If the relationship between the disease process and the visit processes is strong then the disease and visit processes can be modeled jointly as depicted in the expanded state space of [Figure A.1](#) of [Appendix A.2](#); see [Cook and](#)

[Lawless \(2021\)](#) for the corresponding likelihood which could then be penalized. To exploit existing software a more complex EM algorithm would need to be developed but elements of the complete data likelihood would share the same general form as the one presented here. If there are likewise concerns about dependent censoring have an impact on estimators then a joint model could be expanded to include a censoring state. Alternatively if it was thought to be covariate - rather than state-dependent censoring, this could be addressed through use of inverse probability of censoring weights.

Chapter 3

Variable Selection for Illness-death Processes under Dual Observation Schemes

3.1 Introduction

Multistate models can provide important insights into the dynamic nature of complex disease processes and are being used increasingly often in public health research ([Cook and Lawless, 2018](#); [Andersen and Ravn, 2023](#)). In particular, intensity-based regression models can be used to estimate the effect of prognostic factors on the risk of transitions to different disease states and predict disease course ([Touraine and others, 2016](#)). The illness-death model ([Xu and others, 2010](#)) is a canonical multistate model widely used in epidemiology to estimate disease incidence and prevalence rates ([Keiding, 1991](#)), and is commonly adopted in cohort studies for modeling disease progression and complications in

affected individuals. In oncology research, for example, the illness-death model has been used to study the risk of cancer progression, relapse, or the development of new metastases in a framework that recognizes the competing risk of death ([Chae and others, 2022](#)). In transplant research, it has been used to study the risk of graft rejection ([Prentice and others, 1978](#)), and in hepatology it has been used to study the risk of bleeding among individuals with cirrhosis ([Andersen and others, 2000](#)). The illness-death process is also a simple model for considering the joint distribution of the time to a non-fatal event and death. It recognizes that event-free death precludes the future occurrence of the non-fatal event, while the occurrence of the non-fatal (intermediate) event alters the risk of death. We consider the setting of a cohort study in which individuals are recruited initially event-free but are at risk of both a non-fatal event and a terminal event from an illness-death process. For simplicity, we refer to the non-fatal event as *disease progression* and the terminal event as death.

Often, disease progression is not apparent to affected individuals. In cancer studies, for example, disease progression status is recorded at regularly scheduled imaging assessments ([Boruvka and Cook, 2016](#)), so an individual may be progression-free at one assessment and meet the criteria for progression at the next assessment. In other disease settings, such as arthritis, joint damage can only be assessed during radiographic examinations ([Cook and Lawless, 2014](#)) and is determined to have reached a certain severity state following a confirmatory X-ray assessment. In osteoporosis studies, elderly patients with low bone density are at high risk of fractures, but many are asymptomatic and are only detected when bone scans are conducted ([Genant and others, 2008](#)). In such cases, one only knows the state occupied at intermittent assessment times, so progression times are interval-censored ([Sun, 2006](#)). For individuals who are progression-free at their last assessment, their ultimate progression status is unknown since they may or may not have progressed

between the last assessment and censoring or death. We use the term *dual observation scheme* (Boruvka and Cook, 2016) for the setting in which progression times are interval-censored and death times are subject to random right-censoring. Likelihoods for such settings require specification of all intensity functions (Xu and others, 2010).

Much research has taken place in the last 30 years on variable selection algorithms. Traditional model selection approaches such as subset selection involve selecting a model from a list of candidate models. However, it can be unstable as they rely on a “discrete” process where each variable is either included or excluded, making the selected model sensitive to subtle changes in the observed data and potentially reducing predictive accuracy (Tibshirani, 1996). Moreover, subset selection can be computationally intensive as the number of covariates increases. Ridge regression (Hoerl and Kennard, 1970) can improve predictive accuracy by adding an L_2 penalty to the observed-data likelihood that shrinks the coefficient estimates and reduces the variance of estimators. However, because ridge regression does not shrink coefficients to zero, it does not perform variable selection. Penalizing the observed data likelihood by an L_1 penalty, the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) combines the strengths of subset selection and ridge regression in the sense that it can reduce the variance of estimators and select variables by shrinking some coefficients to zero. Use of the LASSO penalty has been applied to generalized linear models using the idea of iteratively reweighted least squares (Friedman and others, 2010) and Cox regression (Tibshirani, 1997). The adaptive LASSO (ALASSO) (Zou, 2006) improves on the LASSO by incorporating weights for coefficients, which ensures that the probability of selecting the correct set of non-zero coefficients approaches one as the sample size increases – this is referred to as the so-called oracle property. Motivated by the need to select dummy variables created for a categorical variable with multiple levels as a group, Yuan and Lin (2006) proposed the group LASSO (GLASSO), but this has seen

wider applications to address variable selection in statistical genetics, where sets of genes may be viewed as part of a network with some common function ([Liu and others, 2014](#)).

The literature is relatively sparse on variable selection with multistate processes under intermittent observation. [Machado and others \(2021\)](#) applied penalization to multistate models with interval-censored data, primarily to fit smoothing splines for the baseline intensities in a setting without covariates. [Li and others \(2025\)](#) considered a progressive 3-state process involving transitions from an initial state to an intermediate state and from the intermediate state to the absorbing state under intermittent observation – covariate selection was carried out using the LASSO and GLASSO penalty functions.

Here we consider the problem of variable selection for the illness-death process under dual observation schemes. Key contributions of this work include a variable selection algorithm that recognizes a) the semi-competing risk problem that progression-free death precludes the future occurrence of progression and b) that deaths can occur after progression. The proposed framework enables one to specify different penalty functions for different transitions, with the option of incorporating a group LASSO penalty to penalize coefficients corresponding to the same covariate across different transitions. This can be attractive in the illness-death context if one wishes to select variables for the two death intensities together with a view that they may be affected by the same covariates. In more general multistate processes, there may be other rationales for grouping coefficients across different types of transitions. We propose the likelihood and expectation-maximization (EM) algorithm ([Dempster and others, 1977](#)) tailored for the dual observation scheme, which is common for problems where the illness-death model naturally applies, such as cancer trials studying risk factors for disease progression or relapse. Another contribution is the introduction of new metrics for effectively evaluating the performance of variable selection algorithms involving complex processes, and enabling meaningful comparison with

established variable selection methods. These are useful for the present problem, and the framework we propose for the illness-death process itself can be generalized to deal with more elaborate processes with more complex state space structures; we comment on this briefly in the discussion.

The remainder of this chapter is organized as follows. In Section 3.2.1, we describe a motivating problem involving the study of age-related cognitive decline among participants in Alzheimer’s disease centers. In Section 3.2.2, we introduce notation, define key functions, and construct the penalized observed data likelihood for variable selection. In Section 3.3, we describe an expectation-maximization (Dempster *and others*, 1977) algorithm which facilitates the use of existing algorithms to maximize the penalized observed data likelihood. Section 3.4 reports the findings from simulation studies designed to evaluate the finite sample performance of the variable selection algorithm, including comparisons with penalized Cox regression applied to the progression-free and overall survival times. In Section 3.5, we apply the proposed method to data from the National Alzheimer’s Coordinating Center (NACC) and examine prognostic factors for the development of dementia jointly with death. Concluding remarks and topics for future work are given in Section 3.6.

3.2 Motivating Problem and General Framework

3.2.1 Prognostic Factors of Dementia in National Alzheimer’s Coordinating Center (NACC) Cohort

The National Alzheimer’s Coordinating Center (NACC) maintains a data repository that collects longitudinal data from 37 Alzheimer’s Disease and Research Centers (ADRCs) across the United States. We utilized the Uniform Data Set (UDS), which has been col-

lected using a standard evaluation protocol and includes all subjects enrolled across ADRCs since 2005, with data frozen as of March 2024. The UDS includes information on participants collected during in-person visits or through phone calls. Participants were scheduled for annual follow-up assessments, but the actual visit times are irregular and vary from person to person. The progression of cognitive and functional decline is measured by CDR[®] Dementia Staging Instrument ([Berg, 1984](#)), a 4-point scale. According to the CDR definition, with a score of 0 representing no impairment, while scores of 0.5, 1, 2, and 3 indicating questionable, mild, moderate, and severe impairment, respectively. A subject is considered as healthy if the CDR is 0 or 0.5, and cognitively impaired if the CDR is 1 or higher. Our interest lies in analyzing aging-related onset of dementia, so we restrict our attention to individuals who entered the study after the age of 50 and were initially healthy. The potential predictors, including race, sex, number of APOE e4 alleles, education, and preliminary assessments of cognitive decline symptoms, are selected by referring to the study by [Pang and others \(2023\)](#), which provides clinical variables associated with the progression of mild cognitive impairment and Alzheimer’s disease. A dictionary of candidate variables, as well as details on how they were coded, is provided in Appendix [B.2](#). After excluding subjects with incomplete covariate information, 6398 individuals from 36 ADRCs remained in the sample for further analysis.

Among the 6398 individuals, 394 (6.2%) were observed to develop cognitive impairment, and the exact time of death was recorded for 504 (7.9%) of those included. The average age of the participants when they entered the study was 73.1 years, and on average, each participant had 2.2 (min = 1, max = 8) assessments. The median duration of follow-up is 5.25 years with the 25th and 75th percentiles being 2.75 and 6.83, respectively (min = 0.083, max = 9.00). We define time as the time from the baseline assessment upon entry to the study. Figure [3.1](#) shows the Kaplan–Meier estimates ([Kaplan and Meier, 1958](#))

of survival probabilities for overall survival and progression-free survival (PFS) with 95% confidence bands. We follow FDA guidelines for cancer clinical trials ([Zeng and others, 2018](#)) and define the PFS time as the time to the first time subjects were known to progress to cognitive impairment or death. The gap between the two curves reflects the proportion of individuals who developed cognitive impairment but remained alive, which started to increase around one year since baseline assessment and then stabilized. Figure 3.1 (b) shows sample paths for four selected subjects reflecting four types of observed data resulted from the dual observation scheme. Subjects 1 and 2 were known to be healthy at the last visit, whereas subjects 3 and 4 were observed to be cognitively impaired. The time to death is right-censored for subjects 1 and 4 due to loss to follow-up, whereas subjects 2 and 3 were observed to die. We aim to identify prognostic factors for cognitive impairment and death.

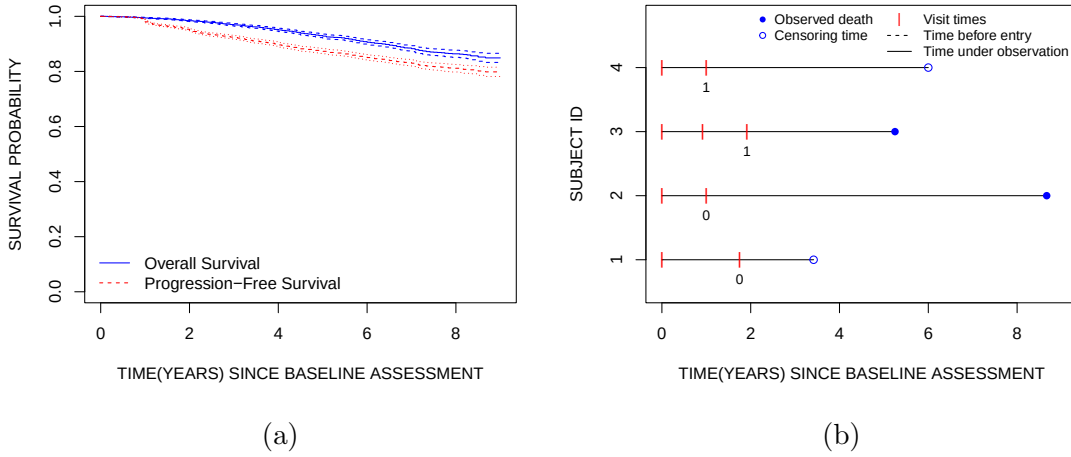


Figure 3.1: (a) Kaplan–Meier estimates of survival probabilities for overall survival and progression-free survival (PFS) with 95% confidence bands; (b) sample paths for four selected subjects with disease status at the last visit, 1 indicating cognitive impairment and 0 being healthy, and death or censoring at the end of study.

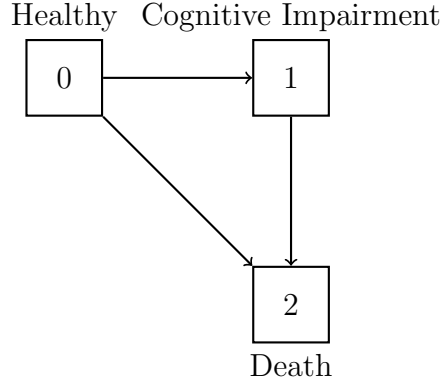


Figure 3.2: A 3-state illness-death model characterizing the onset of cognitive impairment for the samples from the NACC

3.2.2 The Penalized Observed Data Log-likelihood under a Dual Observation Scheme

We consider an illness-death process comprising with state 0 representing the initial healthy state, state 1 as the intermediate state, entered upon the development of some undesirable condition, and state 2 as an absorbing state often representing death. The illness-death process allows for transitions from the initial healthy state to the intermediate illness state, from the healthy state to death directly, and from the intermediate state to death representing death following the intermediate event; see Figure 3.2. This model can be used to characterize the progression of diseases that are hard to cure, for example, cognitive impairment among elderly people or cancer. Let $\{Z(s), s \geq 0\}$ denote the disease process with $Z(t)$ being the state occupied at time t and the state space is $\mathcal{S} = \{0, 1, 2\}$. Assume that $Z(0) = 0$.

Let $H(t) = \{Z(u), 0 \leq u \leq t, \mathbf{w}\}$ denote the complete history of the disease process up to time t along with the fixed $p \times 1$ covariate vector $\mathbf{w} = (w_1, \dots, w_p)'$. Transitions from

state j to state k are governed by the intensity

$$\lambda_{jk}(t \mid H(t^-)) = \lim_{\Delta t \rightarrow 0} \frac{P(Z(t + \Delta t^-) = k \mid Z(t^-) = j, H(t^-))}{\Delta t}, \quad j < k, j, k = 0, 1, 2.$$

Under a modulated Markov model with multiplicative covariate effects, $\lambda_{01}(t \mid H(t^-))$ can be expressed as

$$\lambda_{01}(t \mid \mathbf{w}) = \lambda_{01}(t) \exp(\mathbf{w}' \boldsymbol{\beta}_{01}), \quad (3.1)$$

where $\lambda_{01}(t)$ is the baseline intensity for transitions from state 0 to state 1 and $\boldsymbol{\beta}_{01}$ is an associated vector of regression coefficients. Again, under a modulated Markov model, the intensity for transitions from state j to state 2, $\lambda_{j2}(t \mid H(t^-))$, is given by

$$\lambda_{j2}(t \mid \mathbf{w}) = \lambda_{j2}(t) \exp(\mathbf{w}' \boldsymbol{\beta}_{j2}), \quad (3.2)$$

where $\lambda_{j2}(t)$ is the baseline $j \rightarrow 2$ intensity, and $\boldsymbol{\beta}_{j2}$ is a vector of associated regression coefficients, $j = 0, 1$. We further assume the death baseline intensities are proportional with $\lambda_{12}(t) = \lambda_{02}(t) \exp(\alpha)$, so that the mortality rate is $\exp(\alpha)$ times higher for an individual who has experienced the intermediate event compared to an individual who has not.

We now consider the case where individuals' status with respect to the intermediate event is only known at intermittent times. In this case, the entry time to state 1 may be viewed as interval censored while the state 2 entry time is subject to right-censoring. Let T denote the entry time to state 2, C denote a random right-censoring time, and let $V = \min(T, C)$ and $\Delta = I(T \leq C)$. Let v and δ denote the observed value of V and Δ , respectively. We now consider a sample of n independent processes and introduce the subscript i for individual i , $i = 1, \dots, n$. We let $A_{i1}, \dots, A_{iM_i}$ denote a sequence of M_i random visit times for individual i over $[0, V_i)$. The realized sequence of

visit times is denoted as $a_{i1}, \dots, a_{im_i}$ with $a_{i0} = 0$. The data for individual i is denoted $D_i^\circ = \{(a_{ir}, Z_i(a_{ir})), r = 1, \dots, m_i, v_i, \delta_i, \mathbf{w}_i\}$. The observed history up to a_{ir} ($a_{ir} < v_i$) is $\bar{H}_i^\circ(a_{ir}) = \{(a_{il}, Z_i(a_{il})), l = 1, \dots, r, \mathbf{w}_i\}$. Under non-informative and conditionally independent censoring, the observed data likelihood for an individual with observed data D_i° takes the form

$$\prod_{r=1}^{m_i} P(A_{ir} = a_{ir}, Z_i(a_{ir}) \mid \bar{H}_i^\circ(a_{i,r-1})) \times \left\{ \sum_{j=0}^1 P(A_{i,m_i+1} > v_i^-, Z_i(v_i^-) = j \mid \bar{H}_i^\circ(a_{im_i})) \lambda_{j2}^{\delta_i}(v_i \mid \bar{H}_i^\circ(a_{im_i}), Z_i(v_i^-) = j) \right\}.$$

If the visit process is non-informative it is sufficient to work with the partial likelihood

$$\prod_{r=1}^{m_i} P(Z_i(a_{ir}) \mid A_{ir} = a_{ir}, \bar{H}_i^\circ(a_{i,r-1})) \left\{ \sum_{j=0}^1 P(Z_i(v_i^-) = j \mid \bar{H}_i^\circ(a_{im_i})) \lambda_{j2}^{\delta_i}(v_i \mid \bar{H}_i^\circ(a_{im_i}), Z_i(v_i^-) = j) \right\},$$

and under the assumption of a conditionally independent visit process (CIVP) (Cook and Lawless, 2021), this can be simplified to

$$\prod_{r=1}^{m_i} P(Z_i(a_{ir}) \mid H_i^\circ(a_{i,r-1})) \left\{ \sum_{j=0}^1 P(Z_i(v_i^-) = j \mid H_i^\circ(a_{im_i})) \lambda_{j2}^{\delta_i}(v_i \mid H_i^\circ(a_{im_i}), Z_i(v_i^-) = j) \right\},$$

where $H_i^\circ(a_{ir}) = \{Z_i(a_{il}), l = 1, \dots, r, \mathbf{w}_i\}$ and the likelihood contributions treat the visit times as fixed. Under the modulated Markov assumption, given covariates $\mathbf{w}_i$, the observed log-likelihood for n independent paths can be written as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \left\{ \sum_{r=1}^{m_i} \log P(Z_i(a_{ir}) \mid Z_i(a_{i,r-1}), \mathbf{w}_i) + \log \left[\sum_{j=0}^1 P(Z_i(v_i^-) = j \mid Z_i(a_{im_i}), \mathbf{w}_i) \lambda_{j2}^{\delta_i}(v_i \mid \mathbf{w}_i) \right] \right\}, \quad (3.3)$$

where $\boldsymbol{\theta}$ represents a vector of parameters involved in the illness-death model. We defer the discussion of the specific form of $\boldsymbol{\theta}$ to Section 3.3.3 where we impose assumptions on the baseline intensities. As we point out later, this is a function we need to evaluate in order to calculate the conditional expectations we require for the Expectation-Maximization algorithm (Dempster *and others*, 1977).

To identify important prognostic variables among a large number of candidates, we carry out variable selection by penalizing the observed data likelihood (3.3). The penalized version of (3.3) is written generally as

$$\ell_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) = \frac{1}{n}\ell(\boldsymbol{\theta}) - p_{\boldsymbol{\gamma}}(\boldsymbol{\beta}), \quad (3.4)$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}'_{01}, \boldsymbol{\beta}'_{02}, \boldsymbol{\beta}'_{12})'$, $p_{\boldsymbol{\gamma}}(\boldsymbol{\beta})$ is a penalty function of the regression coefficients in the three intensities, and $\boldsymbol{\gamma}$ is a vector of hyperparameters selected based on criteria such as cross-validation (Stone, 1974), Akaike Information Criterion (AIC) (Akaike, 2003), or Bayesian Information Criterion (BIC) (Schwarz, 1978). The penalty function can take a flexible form, with options including common shrinkage penalties, such as LASSO, ALASSO, and GLASSO or their combinations, depending on the specific objective of variable selection - see Section 3.3.4 for further discussion. A direct maximization of (3.4) requires specialized software and is a challenging high-dimensional optimization problems since it is a function of all parameters in all intensities. The EM algorithm we propose circumvents this by exploiting the partition of the complete data likelihood, reducing the dimension of the optimization problem, and enabling the use of existing variable selection software for generalized linear models. We describe this in the next section.

3.3 A Penalized Expectation-Maximization Algorithm

3.3.1 Likelihood Construction for an Illness-Death Process under Continuous Observation

We adopt a counting process notation to express the complete data log-likelihood and introduce the subscript i to label individuals in a sample of n independent processes. We let $dN_{ijk}(s)$ indicate a $j \rightarrow k$ transition is made at time s for individuals i , $i = 1, \dots, n$. Let $Y_i(s) = I(s \leq C_i)$, $\bar{Y}_{i0}(s) = Y_i(s)I(Z_i(s^-) = 0)$ and $\bar{Y}_{i1}(s) = Y_i(s)I(Z_i(s^-) = 1)$ indicate whether the subject is under observation, at risk of making a transition out of state 0 and out of state 1 at time s , respectively. For a sample of n independent observations generated from the illness-death process, the complete data log-likelihood under a continuous observation scheme takes the form

$$\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{j=0}^1 \sum_{k \in S_j} \int_0^\infty \bar{Y}_{ij}(s) \{dN_{ijk}(s) \log d\Lambda_{ijk}(s | \mathbf{w}_i) - d\Lambda_{ijk}(s | \mathbf{w}_i)\}, \quad (3.5)$$

where S_j is the set of states that can be directly reached through allowable transitions from state j , and $d\Lambda_{ijk}(s | \mathbf{w}_i) = d\Lambda_{jk}(s) \exp(\mathbf{w}_i' \boldsymbol{\beta}_{jk})$ is the subject-specific increment in the cumulative intensity function. For the illness-death model, $S_0 = \{1, 2\}$ and $S_1 = \{2\}$.

We next describe an implementation of the expectation-maximization algorithm ([Dempster and others, 1977](#)) to maximize the penalized observed data log-likelihood in (3.4) using the complete data log-likelihood in (3.5).

3.3.2 The Penalized Expectation-Maximization Algorithm

Expectation of the complete data log-likelihood (the E-step) is taken given the observed data, and the M-step updates the parameters estimated by maximizing the expectation. The two steps are repeated iteratively until the convergence criteria are satisfied. We first assume that $\lambda_{01}(t \mid \mathbf{w})$ does not share any parameters with $\lambda_{j2}(t \mid \mathbf{w})$, $j = 0, 1$. Let $\boldsymbol{\theta}_1$ index $\lambda_{01}(t \mid \mathbf{w})$ and $\boldsymbol{\theta}_2$ index $\lambda_{02}(t \mid \mathbf{w})$ and $\lambda_{12}(t \mid \mathbf{w})$. We then rewrite $\ell_c(\boldsymbol{\theta})$ as

$$\ell_c(\boldsymbol{\theta}) = \ell_{c1}(\boldsymbol{\theta}_1) + \ell_{c2}(\boldsymbol{\theta}_2), \quad (3.6)$$

where

$$\ell_{c1}(\boldsymbol{\theta}_1) = \sum_{i=1}^n \int_0^\infty \bar{Y}_{i0}(s) \{dN_{i01}(s) \log d\Lambda_{i01}(s \mid \mathbf{w}_i) - d\Lambda_{i01}(s \mid \mathbf{w}_i)\} \quad (3.7)$$

contains parameters related to the intensity for transitions into state 1 and

$$\ell_{c2}(\boldsymbol{\theta}_2) = \sum_{i=1}^n \sum_{j=0}^1 \int_0^\infty \bar{Y}_{ij}(s) \{dN_{ij2}(s) \log d\Lambda_{ij2}(s \mid \mathbf{w}_i) - d\Lambda_{ij2}(s \mid \mathbf{w}_i)\} \quad (3.8)$$

contains parameters for the transition intensities into the absorbing state 2.

Let $d\bar{N}_{ijk}(s) = \bar{Y}_{ij}(s)dN_{ijk}(s)$ and $d\bar{N}_i(s) = (d\bar{N}_{i01}(s), d\bar{N}_{i02}(s), d\bar{N}_{i12}(s))'$. Then let $D_i = \{Y_i(s), d\bar{N}_i(s), s > 0, \mathbf{w}_i\}$ denote the complete data, where all transitions are observed subject to right-censoring. Let $\boldsymbol{\theta}^{(r)}$ denote the estimate of $\boldsymbol{\theta}$ at the r th iteration. Given the observed data $\mathbf{D}^\circ = (D_1^\circ, \dots, D_n^\circ)'$ for a sample of n observations, the conditional expectation of (3.6) evaluated at $\boldsymbol{\theta}^{(r)}$ is

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) = Q_1(\boldsymbol{\theta}_1; \boldsymbol{\theta}^{(r)}) + Q_2(\boldsymbol{\theta}_2; \boldsymbol{\theta}^{(r)}), \quad (3.9)$$

where $Q_k(\boldsymbol{\theta}_k; \boldsymbol{\theta}^{(r)}) = E\{\ell_{ck}(\boldsymbol{\theta}_k) \mid \mathbf{D}^\circ; \boldsymbol{\theta}^{(r)}\}$, $k = 1, 2$. These are given by

$$Q_1(\boldsymbol{\theta}_1; \boldsymbol{\theta}^{(r)}) = \sum_{i=1}^n \int_0^\infty \left\{ E[\bar{Y}_{i0}(s) dN_{i01}(s) \mid D_i^\circ; \boldsymbol{\theta}^{(r)}] \log \Lambda_{i01}(s \mid \mathbf{w}_i) \right. \\ \left. - E[\bar{Y}_{i0}(s) \mid D_i^\circ; \boldsymbol{\theta}^{(r)}] d\Lambda_{i01}(s \mid \mathbf{w}_i) \right\} \quad (3.10)$$

and

$$Q_2(\boldsymbol{\theta}_2; \boldsymbol{\theta}^{(r)}) = \sum_{i=1}^n \sum_{j=0}^1 \int_0^\infty \left\{ E[\bar{Y}_{ij}(s) dN_{ij2}(s) \mid D_i^\circ; \boldsymbol{\theta}^{(r)}] \log \Lambda_{ij2}(s \mid \mathbf{w}_i) \right. \\ \left. - E[\bar{Y}_{ij}(s) \mid D_i^\circ; \boldsymbol{\theta}^{(r)}] d\Lambda_{ij2}(s \mid \mathbf{w}_i) \right\}. \quad (3.11)$$

Variable selection is achieved by maximizing a penalized version of (3.9) given by

$$Q_\gamma(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) = \frac{1}{n} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) - p_\gamma(\boldsymbol{\beta}). \quad (3.12)$$

We next discuss an appealing framework for the specification of the intensity functions, facilitating a convenient implementation of the proposed EM algorithm.

3.3.3 A Markov Formulation with Piecewise-Constant Baseline Intensities

Complete Data Log-Likelihood Component for Progression

For $0 \rightarrow 1$ transitions, we adopt a piecewise constant assumption for the baseline intensity $\lambda_{01}(t)$ in (3.1). We let $0 = b_{10} < b_{11} < \dots < b_{1,q_1-1} < b_{1,q_1} = \infty$ denote a sequence of $q_1 + 1$ pre-specified cut-points giving q_1 pieces. Under a piecewise-constant specification of the

baseline intensities, the $0 \rightarrow 1$ baseline intensity can be written as

$$d\Lambda_{01}(s) = \prod_{h=1}^{q_1} \lambda_{1h}^{b_{1h}(s)} ds = \prod_{h=1}^{q_1} \exp(\alpha_{1h} b_{1h}(s)) ds,$$

where $b_{1h}(s) = I(s \in B_{1h})$ with $B_{1h} = (b_{1,h-1}, b_{1h}]$, and $\lambda_{1h} > 0$ is the value for the baseline intensity for $t \in B_{1h}$, and $\alpha_{1h} = \log \lambda_{1h}$, $h = 1, \dots, q_1$.

Let $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12}, \dots, \alpha_{1q_1})'$ be a $q_1 \times 1$ vector and $\boldsymbol{\theta}_1 = (\boldsymbol{\alpha}'_1, \boldsymbol{\beta}'_{01})'$. The complete data partial log-likelihood for transitions from state 0 to state 1 is then

$$\begin{aligned} \ell_{c1}(\boldsymbol{\theta}_1) &= \sum_{i=1}^n \int_0^\infty \bar{Y}_{i0}(s) \left\{ dN_{i01}(s) \left(\sum_{h=1}^{q_1} \alpha_{1h} b_{1h}(s) + \mathbf{w}'_i \boldsymbol{\beta}_{01} \right) - \sum_{h=1}^{q_1} b_{1h}(s) ds \exp(\alpha_{1h} + \mathbf{w}'_i \boldsymbol{\beta}_{01}) \right\} \\ &= \sum_{i=1}^n \sum_{h=1}^{q_1} \{ N_{i01h} (\alpha_{1h} + \mathbf{w}'_i \boldsymbol{\beta}_{01}) - u_{i01h} \exp(\alpha_{1h} + \mathbf{w}'_i \boldsymbol{\beta}_{01}) \}, \end{aligned} \quad (3.13)$$

where $N_{i01h} = \int_0^\infty \bar{Y}_{i0}(s) b_{1h}(s) dN_{i01}(s)$ indicates whether a $0 \rightarrow 1$ transition occurred over B_{1h} in the complete data setting, and $u_{i01h} = \int_0^\infty \bar{Y}_{i0}(s) b_{1h}(s) ds$ records the time at risk for a $0 \rightarrow 1$ transition over B_{1h} while under observation in the complete data setting. Both of these quantities are unknown under the dual observation scheme, so we must take the conditional expectation of them given the observed data.

The conditional expectation of $\ell_{c1}(\boldsymbol{\theta}_1)$ given the observed data and evaluated at $\boldsymbol{\theta}^{(r)}$ is of the form

$$Q_1(\boldsymbol{\theta}_1; \boldsymbol{\theta}^{(r)}) = \sum_{i=1}^n \sum_{h=1}^{q_1} \left\{ \eta_{i01h}^{(r)} (\alpha_{1h} + \mathbf{w}'_i \boldsymbol{\beta}_{01}) - \mu_{i01h}^{(r)} \exp(\alpha_{1h} + \mathbf{w}'_i \boldsymbol{\beta}_{01}) \right\}, \quad (3.14)$$

where $\eta_{i01h}^{(r)} = E(N_{i01h} \mid D_i^\circ, \boldsymbol{\theta}^{(r)})$, and $\mu_{i01h}^{(r)} = E(u_{i01h} \mid D_i^\circ, \boldsymbol{\theta}^{(r)})$. Steps for the calculations of these conditional expectations are provided in Appendix B.1.

Complete Data Log-likelihood Involving Death Intensities

For the failure time process, we could either adopt a parametric model by imposing a piecewise constant assumption for the baseline intensity $\lambda_{j2}(t)$ with arbitrarily many cutpoints, or adopt a semiparametric Cox model with a baseline hazard $\lambda_{j2}(t)$ of unspecified form. Note that although these are two different approaches, the piecewise constant baseline intensity tends to approximate the nonparametric baseline in the Cox model as the number of cut-points increases. Therefore, we regard the failure time model with a piecewise constant baseline as a more general approach and use it to illustrate the proposed EM algorithm in the subsequent discussions.

We let $0 = b_{20} < b_{21} < \dots < b_{2,q_2-1} < b_{2,q_2} = \infty$ denote a sequence of $q_2 + 1$ pre-specified cut-points giving q_2 pieces written in vector form as $(\alpha_{21}, \dots, \alpha_{2q_2})'$, where $\alpha_{2h} = \log \lambda_{2h}$, $h = 1, \dots, q_2$. With $B_{2h} = (b_{2,h-1}, b_{2h}]$ and $b_{2h}(s) = I(s \in B_{2h})$, the baseline intensity for transitions from state 0 to state 2 can be written as

$$d\Lambda_{02}(s) = \prod_{h=1}^{q_2} \lambda_{2h}^{b_{2h}(s)} ds = \prod_{h=1}^{q_2} \exp(\alpha_{2h} b_{2h}(s)) ds.$$

With $\boldsymbol{\alpha}_2 = (\alpha_{21}, \dots, \alpha_{2q_2}, \alpha)'$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\alpha}_2', \boldsymbol{\beta}_2')'$, the complete data partial log-likelihood of the failure time process can be written analogously to (3.13), as

$$\ell_{c2}(\boldsymbol{\theta}_2) = \sum_{i=1}^n \sum_{j=0}^1 \sum_{h=1}^{q_2} \{N_{ij2h} (\alpha_{2h} + \alpha I(j=1) + \mathbf{w}'_i \boldsymbol{\beta}_{j2}) - u_{ij2h} \exp(\alpha_{2h} + \alpha I(j=1) + \mathbf{w}'_i \boldsymbol{\beta}_{j2})\}, \quad (3.15)$$

where $N_{ij2h} = \int_0^\infty \bar{Y}_{ij}(s) b_{2h}(s) dN_{ij2}(s)$, and $u_{ij2h} = \int_0^\infty \bar{Y}_{ij}(s) b_{2h}(s) ds$, $j = 0, 1$. Given the observed data, this gives the conditional expectation of $\ell_{c2}(\boldsymbol{\theta}_2)$ of the form

$$Q_2(\boldsymbol{\theta}_2; \boldsymbol{\theta}^{(r)}) = \sum_{i=1}^n \sum_{j=0}^1 \sum_{h=1}^{q_2} \left\{ \eta_{ij2h}^{(r)} (\alpha_{2h} + \alpha I(j=1) + \mathbf{w}'_i \boldsymbol{\beta}_{j2}) - \mu_{ij2h}^{(r)} \exp(\alpha_{2h} + \alpha I(j=1) + \mathbf{w}'_i \boldsymbol{\beta}_{j2}) \right\}, \quad (3.16)$$

where $\eta_{ij2h}^{(r)} = E(N_{ij2h} \mid D_i^\circ, \boldsymbol{\theta}^{(r)})$, and $\mu_{ij2h}^{(r)} = E(u_{ij2h} \mid D_i^\circ, \boldsymbol{\theta}^{(r)})$, $j = 0, 1$. Details on the calculations of these conditional expectations are given in Appendix B.1.

From (3.14) and (3.16), one can compute (3.12) and so the next step is to discuss the maximization of (3.12) subject to the penalty.

The Maximization Step

Here we show that $Q_\gamma(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)})$ can be reformulated as a penalized Poisson likelihood such that the maximization step with the penalty function can be carried out with existing software. Under the piecewise-constant assumption for both the intermediate event and the death intensities, (3.12) takes the form

$$\begin{aligned} Q_\gamma(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) = & \sum_{i=1}^n \sum_{j=0}^1 \sum_{k \in S_j} \sum_{h=1}^{q_k} \left[\eta_{ijkh}^{(r)} (\log \lambda_{kh} + \alpha I(j=1) + \mathbf{w}'_i \boldsymbol{\beta}_{jk}) \right. \\ & \left. - \mu_{ijkh}^{(r)} \lambda_{kh} \exp(\alpha I(j=1) + \mathbf{w}'_i \boldsymbol{\beta}_{jk}) \right] - p_\gamma(\boldsymbol{\beta}). \end{aligned} \quad (3.17)$$

Note that (3.17) has the form of a penalized Poisson process likelihood. Specifically, for a “pseudo-individual” labeled (i, j, k, h) , $\eta_{ijkh}^{(r)}$ is the count and $\mu_{ijkh}^{(r)}$ is the duration of time at risk. Thus established software can be used for variable selection with Poisson processes via penalized likelihood.

To do so, it requires the creation of a suitable design matrix. Let $\nu_{khr} = I(h = r)$ for $r, h = 1, \dots, q_k$ and $\boldsymbol{\nu}_{kh} = (v_{kh1}, \dots, v_{khq_k})'$, $h = 1, \dots, q_k$, $k = 1, 2$. Then let $\boldsymbol{\nu}_{k1h} = I(k = 1)\boldsymbol{\nu}_{1h}$, $\boldsymbol{\nu}_{k2h} = (I(k = 2)\boldsymbol{\nu}_{2h}, I(j = 1))'$, and $\bar{\boldsymbol{\nu}}_{kh} = (\boldsymbol{\nu}_{k1h}, \boldsymbol{\nu}_{k2h})'$. Finally, let $\boldsymbol{w}_{i01} = I(j = 0, k = 1)\boldsymbol{w}_i$, $\boldsymbol{w}_{i02} = I(j = 0, k = 2)\boldsymbol{w}_i$, $\boldsymbol{w}_{i12} = I(j = 1, k = 2)\boldsymbol{w}_i$, and $\boldsymbol{w}_{ijk} = (\boldsymbol{w}_{i01}', \boldsymbol{w}_{i02}', \boldsymbol{w}_{i12}')'$. Then for pseudo-individual (i, j, k, l) , let $\boldsymbol{x}_{ijkh} = (\bar{\boldsymbol{\nu}}_{kh}', \boldsymbol{w}_{ijk}')'$ enabling us to rewrite (3.17) as

$$Q_{\gamma}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) = \sum_{i=1}^n \sum_{j=0}^1 \sum_{k \in S_j} \sum_{h=1}^{q_k} \left[\eta_{ijkh}^{(r)}(\boldsymbol{x}_{ijkh}'\boldsymbol{\psi}) - \mu_{ijkh}^{(r)} \exp(\boldsymbol{x}_{ijkh}'\boldsymbol{\psi}) \right] - p_{\gamma}(\boldsymbol{\beta}), \quad (3.18)$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}'_1, \boldsymbol{\alpha}'_2)'$, $\boldsymbol{\beta} = (\boldsymbol{\beta}'_{01}, \boldsymbol{\beta}'_{02}, \boldsymbol{\beta}'_{12})'$ and $\boldsymbol{\psi} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ is a reparameterization of $\boldsymbol{\theta}$, where each element takes on values on the real line. This reformulation allows one to create a pseudo data set, where each observation comes from a pseudo-individual labeled (i, j, k, h) , and employ existing R packages for maximization. Options for specification of the penalty function are discussed in the next section.

3.3.4 Specification and Incorporation of Penalty Functions

We have presented a general framework that allows the incorporation of a penalty term into the three-state illness-death process. In this section, we show different ways of applying penalty functions and some examples.

$p_{\gamma}(\boldsymbol{\beta})$ can take any form of existing penalty term for variable selection, such as LASSO (Tibshirani, 1996), ALASSO (Zou, 2006), and SCAD (Fan and Li, 2001). One may be more interested in identifying key covariates associated with the overall event process rather than selecting transition-specific effects, which motivates the use of GLASSO. The form of the GLASSO penalty term can vary depending on the study's objectives. For instance, if the goal is to select important covariates for the entire three-state process, the GLASSO penalty can be defined as

$$p_{\gamma}(\boldsymbol{\beta}) = \gamma \sum_{l=1}^p (|\beta_{01l}| + |\beta_{02l}| + |\beta_{12l}|),$$

which groups the effects across the three transitions. In some cases, one might prefer to penalize the progression and failure processes separately, motivating the use of different penalty terms for each. For this, (3.12) can be re-expressed as

$$Q_{\gamma}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) = \left[\frac{1}{n} Q_1(\boldsymbol{\theta}_1; \boldsymbol{\theta}^{(r)}) - p_{\gamma_1}(\boldsymbol{\beta}_{01}) \right] + \left[\frac{1}{n} Q_2(\boldsymbol{\theta}_2; \boldsymbol{\theta}^{(r)}) - p_{\gamma_2}(\boldsymbol{\beta}_2) \right], \quad (3.19)$$

where $\boldsymbol{\beta}_2 = (\boldsymbol{\beta}'_{02}, \boldsymbol{\beta}'_{12})'$, and $p_{\gamma_1}(\boldsymbol{\beta}_{01})$ and $p_{\gamma_2}(\boldsymbol{\beta}_2)$ are penalty functions of only $\boldsymbol{\beta}_{01}$ and $\boldsymbol{\beta}_2$ indexed by hyperparameters γ_1 and γ_2 , respectively. The functions associated with the two processes can be maximized separately to obtain an updated estimate for $\boldsymbol{\theta}$. One could set $p_{\gamma_1}(\boldsymbol{\beta}_{01}) = \gamma_1 \sum_{l=1}^p |\beta_{01l}|$ and $p_{\gamma_2}(\boldsymbol{\beta}_2) = \gamma_2 \sum_{l=1}^p \sqrt{\beta_{02l}^2 + \beta_{12l}^2}$ which groups effects on the two transitions to failure. Similarly, if the goal is to identify important variables for progression-free survival, we could group the effects on the 0–1 and 0–2 transitions associated with the same covariate.

Note that the same type of penalty terms can be applied to the two processes. For example, we could apply two LASSO penalties by specifying $p_{\gamma_1}(\boldsymbol{\beta}_{01}) = \gamma_1 \sum_{l=1}^p |\beta_{01l}|$ and

$p_{\gamma_2}(\beta_2) = \gamma_2 \sum_{l=1}^p (|\beta_{02l}| + |\beta_{12l}|)$. One could also set $\gamma = \gamma_1 = \gamma_2$, which is equivalent to applying a single LASSO penalty applied to both the 0-1 process and the failure process using a common tuning parameter γ . With two LASSO penalties, the model allows different tuning parameters for the progression process and the failure process, and implementation requires a two-dimensional grid search, whereas a common tuning parameter reduces it to a one-dimensional search. Performance comparison between these two options is examined in Appendix B.3.

Here the optimal tuning parameter γ is selected using the Bayesian Information Criterion (BIC), defined as

$$\text{BIC}_{\gamma} = -2\ell(\hat{\theta}_{\gamma}) + |\hat{\theta}_{\gamma}| \log n,$$

where $|\hat{\theta}_{\gamma}|$ is the number of non-zero estimated coefficients, n is the number of subjects in the sample, and $\hat{\theta}_{\gamma}$ represents the parameter estimates for the 3-state model with respect to γ . The dimension of grid searching depends on the dimension of γ .

Note that our goal is not to exhaustively cover all possible cases, but rather to illustrate the potential for:

1. Applying existing penalty functions to the joint process,
2. Applying different penalties to the 0–1 process and the failure process separately, and
3. Applying the same type of penalty functions to the 0–1 process and the failure process while allowing for more than one tuning parameter

in the proposed framework. We explore a few representative cases and evaluate their performance through simulation studies in Section 3.4.

3.4 Simulation Studies

Here we conduct simulation studies to assess the performance of the proposed penalized EM algorithm with several penalty functions and compare the performance with alternative methods that might be considered in applications.

3.4.1 Simulation Design

We consider $n = 1000$ independent individuals and generate $n_{sim} = 100$ samples for each of the following parameter settings. We let $\mathbf{w}_i$ be a $p \times 1$ covariate following a multivariate normal distribution $MVN_p(\mathbf{0}, \Sigma)$ with $p = 100$, where Σ is a covariate matrix with an autoregressive dependence structure wherein the (j, k) entry is $\Sigma_{jk} = \kappa^{|j-k|}$, $j, k = 1, \dots, 100$ with $\kappa = 0.5$.

Three distinct scenarios are considered and they differ in how covariates influence the disease process. Table 3.1 gives the specific parameter values. In Setting 1, covariates $w_1, \dots, w_5$ increase the intensities for all three types of transitions by a factor of 1.2 for each one-unit increase in the covariate value. Covariates $w_{51}, \dots, w_{55}$ increase the rate for $0 \rightarrow 1$ transitions by the same factor but do not affect transition intensities for entry into the absorbing state. Finally covariates $w_{96}, \dots, w_{100}$ impact the $1 \rightarrow 2$ transition intensities only. The remaining 85 variables have no effects and hence can be viewed as noise reflecting the sparsity condition for which penalization methods are suited in analyses oriented towards discovery. In Setting 2, a one unit increase in covariates $w_1, \dots, w_5$ increases the intensity for $0 \rightarrow 2$ transitions but not others, while other covariates have the same effects as in Setting 1. In Setting 3, covariates $w_1, \dots, w_5$ reduce the intensity for $0 \rightarrow 1$ and $1 \rightarrow 2$ transitions with relative risks of 0.8, while increasing the intensity for $0 \rightarrow 2$ transitions; the effects of other covariates are the same as in Settings 1 and 2.

Setting	j	k	$\beta_{jk1}, \dots, \beta_{jk5}$	...	$\beta_{jk51}, \dots, \beta_{jk,55}$	...	$\beta_{jk,96}, \dots, \beta_{jk,100}$	λ_{jk}
1	0	1	$\log(1.2)$	0	$\log(1.2)$	0	0	2.54
	0	2	$\log(1.2)$	0	0	0	0	2.56
	1	2	$\log(1.2)$	0	0	0	$\log(1.2)$	3.84
2	0	1	0	0	$\log(1.2)$	0	0	2.21
	0	2	$\log(1.2)$	0	0	0	0	2.11
	1	2	0	0	0	0	$\log(1.2)$	3.17
3	0	1	$\log(0.8)$	0	$\log(1.2)$	0	0	1.70
	0	2	$\log(1.2)$	0	0	0	0	1.75
	1	2	$\log(0.8)$	0	0	0	$\log(1.2)$	2.63

Table 3.1: Parameter values used for Settings 1, 2, and 3 in simulation studies

We set the follow-up period to $(0, \tau]$ with the administrative censoring time $\tau = 1$. Time homogeneous baseline intensities are denoted by λ_{jk} for $j, k = 0, 1, 2$, and we set $\alpha = \log(\lambda_{12}/\lambda_{02}) > 0$ reflect a possible effect of the intermediate event (entry to state 1) on the risk of failure. Specifically, we set $\alpha = \log 1.5$ such that the intensity for death is elevated by 50% if subjects move to state 1 compared to staying at state 0. We generate the paths for each individual following the steps in Appendix A of (Cook and Lawless, 2018). First, times T_{01}^* and T_{02}^* are generated independently by treating the intensity functions $\lambda_{0k}(t | \mathbf{w})$, $k = 1, 2$ as respective hazard functions. Then if $T_{01}^* > T_{02}^*$, the subject makes a transition to state 2 directly, and we record the path taken as $0 \rightarrow 2$, set the actual $0 \rightarrow 2$ transition time as $T_{02} = T_{02}^*$ and the state 2 entry time as $T = T_{02}$. If $T_{01}^* < T_{02}^*$, a transition to state 1 occurs at $T_{01} = T_{01}^*$ and the individual becomes at risk of death from state 1. In such cases, the state 2 entry time T_{12} is generated as a failure time from a distribution with hazard function $\lambda_{12}(t | \mathbf{w})$ and left truncation time T_{01} , and the state 2 entry time is $T = T_{12}$. Before generating the paths, the value of the baseline intensities λ_{01} and λ_{02} are determined through Monte-Carlo approximation based on a large sample

of 10000 subjects to ensure $E[P(T < \tau \mid \mathbf{W}; \boldsymbol{\theta})] = 0.9$ and $E[P(T_{01}^* < T_{02}^* \mid \mathbf{W}; \boldsymbol{\theta})] = 0.5$, which corresponds to an administrative right-censoring rate of 0.1 for the state 2 entry time and there is a 50% probability of passing through state 1, respectively.

The visit process is governed by a time-homogeneous Poisson process adopted with rate $\rho = 8$ or 16. To generate the visit times, we first generate the number of visits $M_i^* \sim \text{Poisson}(\rho)$ for individual i and given m_i^* we generate the times $a_{i1}, a_{i2}, \dots, a_{im_i^*}$ uniformly over $[0, 1]$. We record $V_i = \min(T_i, \tau)$ and $\Delta_i = I(V_i = T_i)$ and retain $m_i < m_i^*$ visits such that $a_{ij} < V_i$, $j = 0, 1, \dots, m_i$. The observed data for subject i is $D_i^\circ = \{(a_{ir}, Z_i(a_{ir})), r = 1, \dots, m_i, v_i, \delta_i, \mathbf{w}_i\}$, with $a_{i0} = 0$ and $Z_i(a_{i0}) = 0$ for $i = 1, 2, \dots, n$.

The initial value for the EM algorithm is denoted by $\boldsymbol{\theta}^{(0)}$ and obtained by setting $\boldsymbol{\beta} = \mathbf{0}$ and fitting a multistate model with time-homogeneous intensities and no covariates via a maximum likelihood. The R function `msm` by [Jackson \(2011\)](#) can provide these. The convergence criterion for the EM algorithm is set to $\max_j |(\boldsymbol{\theta}_j^{(r+1)} - \boldsymbol{\theta}_j^{(r)})/\boldsymbol{\theta}_j^{(r)}| < 10^{-4}$, where $\boldsymbol{\theta}_j^{(r)}$ is the j th element of $\boldsymbol{\theta}$ at the r th iteration. We iterate between the E-step and the M-step until this convergence criterion is satisfied.

We explore two reference methods. The first method disregards information on progression and focuses solely on the failure time, referred to here as the overall survival time. This may be considered if entry time to state 2 is of sole interest. Since the entry time to state 2 is subject only to right-censoring in our settings, we consider penalized Cox regression for the overall survival time accommodating right censoring ([Tibshirani, 1997](#)). The best model is selected using a BIC-related criterion ([Volinsky and Raftery, 2000](#)) given by

$$\text{BIC}_\gamma = -2\ell_{\text{partial}}(\hat{\boldsymbol{\beta}}_\gamma) + |\hat{\boldsymbol{\beta}}_\gamma| \log n_{\text{Cox}},$$

where $\hat{\boldsymbol{\beta}}_\gamma$ is the estimates obtained from the penalized Cox model with the penalty $p_\gamma(\boldsymbol{\beta})$,

$\ell_{\text{partial}}(\widehat{\beta}_{\gamma})$ is the partial log-likelihood of the Cox model evaluated at $\widehat{\beta}_{\gamma}$, and n_{Cox} is the number of observed transitions to state 2. The second method focuses on the time to transition out of state 0, referred to here as the progression-free survival time. Since the progression time is under intermittent observation, yielding an interval $(L, R]$, where L is the last time the subject is observed in state 0, and R is the first time the subject is observed in a state other than state 0, we consider the common ad-hoc approach, right-endpoint imputation, which is adopted if R is finite, with the left endpoint used as a censoring time otherwise. For the Cox model for progression-free survival, we use the same BIC criterion with n_{Cox} defined as the total number of observed transitions to state 1 or state 2.

The results for further investigations into the role of different methods for estimating baseline intensities with right-censored data are provided in the Appendix B.3. The results of using different penalty functions are also included for reference.

3.4.2 Evaluation Metrics

Due to the complexity of the illness-death process and the various roles that covariates may have, we consider several performance evaluation metrics.

Definitions of True Positive Selection Rates

We first consider performance metrics based on the multistate analysis. Let $\mathcal{C}_{jk} = \{l : \beta_{jkl} \neq 0\}$ be the set of covariates which have effects on the $j \rightarrow k$ transition intensity, $\mathcal{D}_{jk} = \{l : \widehat{\beta}_{jkl} \neq 0\}$ be the set of covariates detected to have an effect on the $j \rightarrow k$ transition intensity, and $\mathcal{B}_{jk} = \mathcal{C}_{jk} \cap \mathcal{D}_{jk}$ be the set of covariates in both sets (i.e., the set of covariates that are correctly selected). The true positive rate for covariates on the $j \rightarrow k$ transition intensity is $\text{TP}_{jk} = |\mathcal{B}_{jk}| / |\mathcal{C}_{jk}|$. The overall true positive rate for covariate

selection is defined as the proportion of covariates with one or more non-zero coefficients across all transitions that are selected. With $\mathcal{C} = \mathcal{C}_{01} \cup \mathcal{C}_{02} \cup \mathcal{C}_{12}$ the set of all covariates impacting the illness-death process, $\mathcal{D} = \mathcal{D}_{01} \cup \mathcal{D}_{02} \cup \mathcal{D}_{12}$ the set of all covariates found to have an effect, we define $\mathcal{B} = \mathcal{C} \cap \mathcal{D}$ and define $\text{TP}_{\text{cov}} = |\mathcal{B}| / |\mathcal{C}|$. Performance can also be assessed in terms of the number of non-zero coefficients detected to be non-zero. Here we define $\text{TP}_{\text{coef}} = \sum_{j < k} |\mathcal{B}_{jk}| / \sum_{j < k} |\mathcal{C}_{jk}|$. Performance for identification of covariates for failure (entry to state 2) can focus on the detection of covariates affecting the $0 \rightarrow 2$ and $1 \rightarrow 2$ intensities. For this, we let $\mathcal{C}_2 = \mathcal{C}_{02} \cup \mathcal{C}_{12}$ denote the set of such covariates, $\mathcal{D}_2 = \mathcal{D}_{02} \cup \mathcal{D}_{12}$ denote the set of covariates found to have an effect on the respective intensities, and $\mathcal{B}_2 = \mathcal{C}_2 \cap \mathcal{D}_2$. We then define $\text{TP}_2 = |\mathcal{B}_2| / |\mathcal{C}_2|$. If interest lies in variables impacting progression-free survival, we consider $\mathcal{C}_0 = \mathcal{C}_{01} \cup \mathcal{C}_{02}$ as the set of covariates having an effect, $\mathcal{D}_0 = \mathcal{D}_{01} \cup \mathcal{D}_{02}$ as the set of covariates found to have an effect, and define $\mathcal{B}_0 = \mathcal{C}_0 \cap \mathcal{D}_0$ to give $\text{TP}_0 = |\mathcal{B}_0| / |\mathcal{C}_0|$.

We now consider performance metrics for common methods of analysis, which focus on marginal features such as the overall survival time or the progression-free survival time. Let $\mathcal{D}_2^{\text{Cox}} = \{l : \hat{\beta}_l \neq 0\}$, where $\hat{\beta}_l$ is the estimate from Cox regression; the subscript 2 is to denote that the event being modeled is the entry time to state 2. A covariate may affect overall survival directly through effects on the $0 \rightarrow 2$ or $1 \rightarrow 2$ transition intensities and if interest lies in these we consider covariates in $\mathcal{C}_2$ as the target set of covariates; we let $\mathcal{B}_2^{\text{Cox}} = \mathcal{C}_2 \cap \mathcal{D}_2^{\text{Cox}}$ and we define $\text{TP}_2^{\text{Cox}} = |\mathcal{B}_2^{\text{Cox}}| / |\mathcal{C}_2|$. One could also consider variables that affect any transitions as these will generally impact the overall survival – in this case, interest lies in the full set $\mathcal{C}$. For this, we consider $\mathcal{B}^{\text{Cox}} = \mathcal{C} \cap \mathcal{D}_2^{\text{Cox}}$ and $\text{TP}^{\text{Cox}} = |\mathcal{B}^{\text{Cox}}| / |\mathcal{C}|$; this represents the proportion of covariates having any effect on the illness-death process that are selected by penalized Cox regression for the failure time. When a Cox regression model is used for the progression-free survival analysis, we let

$\mathcal{D}_0^{\text{Cox}} = \{l : \widehat{\beta}_l \neq 0\}$, where $\widehat{\beta}_l$ is the estimate from Cox model for progression-free survival, $\mathcal{B}_0^{\text{Cox}} = \mathcal{C}_0 \cap \mathcal{D}_0^{\text{Cox}}$ and define $\text{TP}_0^{\text{Cox}} = |\mathcal{B}_0^{\text{Cox}}| / |\mathcal{C}_0|$ as the proportion of covariates impacting the sojourn time distribution in state 0 that are identified from a penalized progression-free Cox regression analysis.

Definitions of False Positive Selection Rates

Let $\bar{\mathcal{C}}_{jk} = \{l : \beta_{jkl} = 0\}$ be the set of covariates having no effect on the $j \rightarrow k$ transition intensity and $\mathcal{A}_{jk} = \bar{\mathcal{C}}_{jk} \cap D_{jk}$ be the set of covariates incorrectly found to have an effect. The false positive rate for $j \rightarrow k$ transitions is $\text{FP}_{jk} = |\mathcal{A}_{jk}| / |\bar{\mathcal{C}}_{jk}|$. The complement of $\mathcal{C}$ is the set $\bar{\mathcal{C}}$ of covariates that do not affect the full process (i.e., no element of $\bar{\mathcal{C}}$ affect any intensity). If $\mathcal{A} = \bar{\mathcal{C}} \cap \mathcal{D}$, the overall false positive rate is then $\text{FP}_{\text{cov}} = |\mathcal{A}| / |\bar{\mathcal{C}}|$. We can also define the false positive rate of selecting coefficients as non-zero when they are in fact zero. For this we define $\text{FP}_{\text{coef}} = \sum_{j < k} |\mathcal{A}_{jk}| / \sum_{j < k} |\bar{\mathcal{C}}_{jk}|$. We also let $\bar{\mathcal{C}}_2$ be the set of covariates with no effect on the intensities for entry to state 2; letting $\mathcal{A}_2 = \bar{\mathcal{C}}_2 \cap \mathcal{D}_2$ denote the number of such variables selected, we obtain $\text{FP}_2 = |\mathcal{A}_2| / |\bar{\mathcal{C}}_2|$. Likewise as $\bar{\mathcal{C}}_0$ is the set of variables that do not affect the sojourn time in state 0, we let $\mathcal{A}_0 = \bar{\mathcal{C}}_0 \cap \mathcal{D}_0$ and define $\text{FP}_0 = |\mathcal{A}_0| / |\bar{\mathcal{C}}_0|$ as the false positive rate for the proposed algorithm. For the time to event analyses, we let $\mathcal{A}_2^{\text{Cox}} = \bar{\mathcal{C}}_2 \cap \mathcal{D}_2^{\text{Cox}}$ and define $\text{FP}_2^{\text{Cox}} = |\mathcal{A}_2^{\text{Cox}}| / |\bar{\mathcal{C}}_2|$; similarly we can define $\mathcal{A}^{\text{Cox}} = \bar{\mathcal{C}} \cap \mathcal{D}_2^{\text{Cox}}$ and $\text{FP}^{\text{Cox}} = |\mathcal{A}^{\text{Cox}}| / |\bar{\mathcal{C}}|$. Finally for Cox regression based on progression-free survival, we let $\mathcal{A}_0^{\text{Cox}} = \bar{\mathcal{C}}_0 \cap \mathcal{D}_0^{\text{Cox}}$ and $\text{FP}_0^{\text{Cox}} = |\mathcal{A}_0^{\text{Cox}}| / |\bar{\mathcal{C}}_0|$.

3.4.3 Empirical Results

Tables 3.2 and 3.3 summarize the true positive and false positive rates for the 3-state model compared to the two reference methods based on penalized Cox regression with $\rho = 16$

and $\rho = 8$ governing the visit times, respectively.

One reference method is based on a progression-free survival analysis in which the response is the time spent in state 0 while the other is for overall survival where the response is the time of entry to state 2 - in both cases, the times are treated as right-censored. A progression-free survival analysis is challenging when progression status is only determined at the intermittent observation times. For this, we adopt the convention of using the time individuals are first known to have left the initial state; this can be the time they were detected to have progressed into state 1 or the time they were known to have entered state 2. This follows FDA guidance on how progression-free survival times should be defined in cancer clinical trials ([Zeng and others, 2018](#)) but it does presume that progression did not occur in individuals dying following a final assessment in which they were in the initial state.

We explore three approaches to penalization: (1) a single LASSO penalty applied to all transitions, (2) two separate LASSO penalties with one for the disease progression process and the other for the death process, and (3) a combination of LASSO for the 0-1 transition and GLASSO for transitions to death. Based on the results, we observe that the three-state models generally perform well with a high true positive rate combined with low false positive rates. This holds across all three settings when there is a high frequency of visits ($\rho = 16$). We observe only minor differences in variable selection performance between using a single LASSO penalty and two LASSO penalties across all three settings. When using the LASSO+GLASSO, the coefficient-wise true positive rate (TP_{coef}) tends to be higher for all transitions in Setting 1 and 3. However, in Setting 2, the LASSO+GLASSO approach leads to a lower TP rate for the 0-2 and 1-2 transitions with $TP_{02} = 89.8\%$ and $TP_{12} = 89\%$, respectively, compared to the single LASSO and two LASSO approaches which achieve around 95% in both transition-specific TPs. This results in lower values

Penalization Method	Metric	LASSO	True Positive Rate			False Positive Rate		
			LASSO ¹	LASSO+GLASSO ²		LASSO	LASSO ¹	LASSO+GLASSO ²
SETTING 1								
Multistate-EM	TP_{01}	94.70	94.50	95.70	FP_{01}	1.19	2.13	1.93
	TP_{02}	97.80	97.20	99.20	FP_{02}	0.75	0.76	5.36
	TP_{12}	92.40	92.30	94.10	FP_{12}	2.22	2.21	0.71
	TP_{cov}	97.55	96.60	95.33	FP_{cov}	4.24	4.94	2.56
	TP_{coef}	94.40	94.16	95.76	FP_{coef}	1.37	1.68	2.75
	TP_2	96.90	96.80	94.10	FP_2	2.97	2.97	0.71
	TP_0	98.40	97.60	98.30	FP_0	2.04	2.88	7.39
Cox-OS	TP_2^{cox}	—	83.90	—	FP_2^{cox}	—	1.27	—
	TP^{cox}	—	68.10	—	FP^{cox}	—	1.14	—
Cox-PFS	TP_0^{cox}	—	90.87	—	FP_0^{cox}	—	0.65	—
SETTING 2								
Multistate-EM	TP_{01}	98.00	96.40	96.40	FP_{01}	1.08	0.98	1.08
	TP_{02}	95.80	95.80	89.80	FP_{02}	0.62	0.62	5.36
	TP_{12}	94.40	94.40	89.00	FP_{12}	1.69	1.71	5.40
	TP_{cov}	96.20	95.73	91.80	FP_{cov}	3.39	3.28	1.82
	TP_{coef}	96.07	95.53	91.73	FP_{coef}	1.13	1.10	3.95
	TP_2	95.20	95.20	89.40	FP_2	2.29	2.30	0.71
	TP_0	97.00	96.30	93.20	FP_0	1.70	1.59	6.67
Cox-OS	TP_2^{cox}	—	75.50	—	FP_2^{cox}	—	0.91	—
	TP^{cox}	—	51.60	—	FP^{cox}	—	0.74	—
Cox-PFS	TP_0^{cox}	—	82.10	—	FP_0^{cox}	—	0.76	—
SETTING 3								
Multistate-EM	TP_{01}	95.20	95.90	96.17	FP_{01}	1.03	3.08	3.31
	TP_{02}	88.20	88.60	95.67	FP_{02}	1.00	0.87	5.84
	TP_{12}	91.40	91.50	91.17	FP_{12}	2.18	1.99	1.35
	TP_{cov}	96.50	95.67	94.00	FP_{cov}	4.28	5.73	4.45
	TP_{coef}	92.28	92.68	94.76	FP_{coef}	1.40	1.96	3.46
	TP_2	94.40	94.60	91.17	FP_2	3.19	2.88	1.35
	TP_0	98.00	97.60	97.50	FP_0	2.07	3.88	9.04
Cox-OS	TP_2^{cox}	—	14.60	—	FP_2^{cox}	—	0.20	—
	TP^{cox}	—	7.55	—	FP^{cox}	—	0.19	—
Cox-PFS	TP_0^{cox}	—	20.27	—	FP_0^{cox}	—	0.19	—

¹An alternative LASSO with two tuning parameters, each for the $0 \rightarrow 1$ transition and the death process, respectively.

²LASSO for the $0 \rightarrow 1$ process and GLASSO for the $0 \rightarrow 2$ and $1 \rightarrow 2$ process .

Table 3.2: Empirical results for multistate analysis under independent intermittent observation scheme summarizing true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression for PFS and OS time; $n = 1000$ with $\rho = 16$ as the mean number of visits over $(0, 1]$.

Penalization Method	Metric	LASSO	True Positive Rate			False Positive Rate		
			LASSO ¹	LASSO+GLASSO ²		LASSO	LASSO ¹	LASSO+GLASSO ²
SETTING 1								
Multistate-EM	TP_{01}	68.60	68.50	70.00	FP_{01}	0.53	1.43	1.26
	TP_{02}	84.20	83.40	85.00	FP_{02}	0.57	0.51	4.36
	TP_{12}	71.90	72.30	77.40	FP_{12}	2.34	2.84	1.81
	TP_{cov}	81.10	78.87	76.47	FP_{cov}	2.34	2.84	1.81
	TP_{coef}	73.04	73.00	76.00	FP_{coef}	0.85	1.07	2.14
	TP_2	81.10	81.00	77.40	FP_2	1.97	1.72	0.72
	TP_0	71.91	72.30	77.40	FP_0	1.13	1.91	5.59
Cox-OS	TP_2^{cox}	–	83.90	–	FP_2^{cox}	–	1.27	–
	TP^{cox}	–	68.10	–	FP^{cox}	–	1.14	–
Cox-PFS	TP_0^{cox}	–	87.93	–	FP_0^{cox}	–	0.84	–
SETTING 2								
Multistate-EM	TP_{01}	77.37	76.40	76.60	FP_{01}	0.51	0.74	0.71
	TP_{02}	78.99	78.20	71.00	FP_{02}	0.31	0.29	4.18
	TP_{12}	80.20	79.40	71.20	FP_{12}	0.93	0.91	4.17
	TP_{cov}	78.99	78.20	73.20	FP_{cov}	1.60	1.76	1.11
	TP_{coef}	78.86	78.00	71.60	FP_{coef}	0.58	0.65	3.04
	TP_2	79.70	78.90	71.10	FP_2	1.17	1.13	0.46
	TP_0	78.18	77.30	73.90	FP_0	0.82	1.03	5.03
Cox-OS	TP_2^{cox}	–	75.50	–	FP_2^{cox}	–	0.91	–
	TP^{cox}	–	51.60	–	FP^{cox}	–	0.74	–
Cox-PFS	TP_0^{cox}	–	77.40	–	FP_0^{cox}	–	0.56	–
SETTING 3								
Multistate-EM	TP_{01}	69.39	74.30	71.17	FP_{01}	1.25	2.92	2.37
	TP_{02}	62.63	60.40	63.67	FP_{02}	1.61	0.95	2.65
	TP_{12}	67.17	64.50	62.50	FP_{12}	2.87	1.74	0.96
	TP_{cov}	74.80	74.00	69.56	FP_{cov}	2.87	1.74	0.96
	TP_{coef}	67.15	67.60	67.96	FP_{coef}	1.90	1.85	2.61
	TP_2	72.63	69.00	62.50	FP_2	4.39	2.63	0.96
	TP_0	73.47	76.00	73.33	FP_0	2.90	3.79	6.44
Cox-OS	TP_2^{cox}	–	14.60	–	FP_2^{cox}	–	0.20	–
	TP^{cox}	–	7.55	–	FP^{cox}	–	0.19	–
Cox-PFS	TP_0^{cox}	–	16.00	–	FP_0^{cox}	–	0.12	–

¹An alternative LASSO with two tuning parameters, each for the $0 \rightarrow 1$ transition and the death process, respectively.

²LASSO for the $0 \rightarrow 1$ process and GLASSO for the $0 \rightarrow 2$ and $1 \rightarrow 2$ process.

Table 3.3: Empirical results for multistate analysis under independent intermittent observation scheme summarizing true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression for PFS and OS time; $n = 1000$ with $\rho = 8$ as the mean number of visits over $(0, 1]$.

of TP_{cov} and TP_{coef} for the LASSO+GLASSO approach as well. There is also a higher FP_{02} across all three settings - this can be explained by how GLASSO groups covariate effects on the transitions to the death state. If the effects for both transitions are large (e.g., in Setting 1 and Setting 3), grouping them together strengthens the signal, resulting in a higher probability of selection compared to when the variables are not grouped. On the other hand, if some covariates affect only one transition to death, grouping them can cause a decrease in the chance of selection and an increase in false positive selection for the transitions where there is no effect. This can be seen in Setting 2, where the TP_{02} and TP_{12} rates degrade when GLASSO is used for the death process instead of LASSO.

The two Cox models fail to capture certain aspects of the three-state model and yield less satisfactory variable selection results, whereas the three-state model maintains good performance across all three settings. The Cox regression model faces two major challenges. First, by collapsing certain states and focusing on a two-state failure process, they are not equipped to assess covariate effects on those transitions. For example, since the PFS analysis merges states 1 and 2, it cannot evaluate covariate effects on the 1-2 transition intensity. Following a similar logic, Cox-OS is not equipped to detect covariate effects on the 0-1 transition intensity because it collapses states 0 and 1 into a single “alive” state. Another challenge with the Cox model is that by collapsing two of the three states and focusing solely on a two-state failure process, it fails to account for the distinct underlying transition paths between the states. As a result, the estimated effects obtained from the Cox model represent a weighted average of the effects on the two possible transitions under this misspecified model. If a strong effect exists on one transition but not on the other, the estimated effects in the Cox model tend to be attenuated and the weakened signals would ultimately result in compromised variable selection performance. For example, Cox-OS yields TP_2^{cox} of only 83.9% (compared to 96.8% for the multistate model) and an overall

TP^{cox} of just 68.2% (compared to 94.6% for the multistate model). These discrepancies are even more pronounced in Setting 2 due to the lack of strong effects for the first five covariates on the $1 \rightarrow 2$ transitions. Similarly, Cox-PFS performs worse than the multistate model with TP_0^{cox} of 90.87% in Setting 1 and 82.1% in Setting 2, whereas the corresponding TP_0 values for the multistate model are 97.6% and 96.3%, respectively. A more extreme case is when the effects on different transitions act in opposite directions, where they may cancel each other out under a Cox model and result in an estimated effect close to zero. This extreme case is considered in Setting 3, where the first five covariates are assigned positive effects on the $0 \rightarrow 2$ transition but negative effects on both the $0 \rightarrow 1$ and $1 \rightarrow 2$ transitions. As a result, Cox-OS identifies only 14.6% of the covariates that truly affect the risk of death, and this number drops to 7.55% when taking those covariates that also influence the $0 \rightarrow 1$ transition as well. Similarly, Cox-PFS achieves a true positive rate of only 20.27%.

When $\rho = 8$, the right-censoring rate remains fixed while the expected frequency of intermittent visits changes. Therefore, changes in visit frequency only affect the performance of the 3-state model and the PFS analysis, whereas the performance of the Cox model for overall survival remains the same. Comparing scenarios with higher visit frequencies to those with lower frequencies, the 3-state model shows a decline in all true positive rates. This results in some slight advantage for the Cox model over the 3-state model in Setting 1 when it comes to identifying covariates that affect risk of death ($TP_2^{\text{cox}} = 83.90\%$ where as $TP_2 = 81\%$), and Cox-PFS outperforms the multistate model with $TP_0^{\text{cox}} = 87.93\%$ while $TP_0 = 72.3\%$. However, in Settings 2 and 3, the 3-state model continues to show its strengths in identifying covariates important to death or covariates important to progression-free survival time.

3.5 Selecting Predictors for Dementia Onset and Death

We applied the proposed penalized EM algorithm to select predictors for dementia onset and death among subjects in the National Alzheimer’s Coordinating Center (NACC) cohort. Specifically, we applied the algorithm under an illness-death model and illustrated the use of three different penalization approaches. Additionally, we included Cox models for progression-free survival (PFS) and overall survival (OS) as reference methods, which are commonly used in practice when interests lie in disease-free survival and the risk of death, respectively. For the illness-death model, the cut-points are set as 3 and 6 for the piecewise constant baseline for all transitions. When dealing with heavily censored data, using the traditional BIC can lead to overly stringent model selection. To address this, we applied a sample-adjusted BIC (SABIC) (Chaurasia, 2023) for the illustrative data analysis. We found that using a modified BIC for smaller sample sizes is effective in practice when the censoring rate is high. The performance of variable selection using SABIC in progressive multistate models was examined and presented in Li and others (2025) and it generally shows good performance in variable selection.

To illustrate how the optimal tuning parameter(s) are selected, we conducted a one-dimensional search in the application of a single LASSO and a two-dimensional grid search in the use of two LASSO penalties. Figure 3.3(a) shows the change in the SABIC over a sequence of values of $\log \gamma$ when using a single LASSO penalty and the optimal model minimizes the SABIC. Figure 3.3(b) is a heat map of the SABIC values that change with the values of $\log(\gamma_{01})$ (tuning parameter for the 0-1 process) and $\log \lambda_{death}$ (tuning for the death process). The optimal tuning parameters are $(\gamma_{01}, \gamma_{death}) = (\exp(-5.4), \exp(-5.8))$ which minimizes the SABIC over the searching grid. Illustrating how the estimates shrink over a two-dimensional grid is not straightforward and can make the plot complex. However,

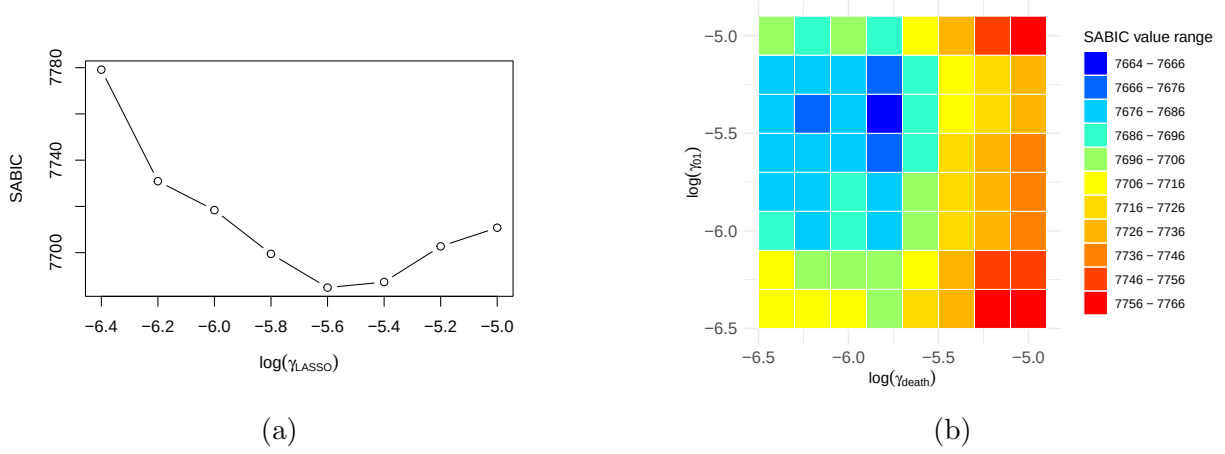


Figure 3.3: (a) Changes in SABIC with respect to $\log(\gamma_{\text{LASSO}})$ using LASSO penalty and (b) heatmap of SABIC with respect to two tuning parameters $\log \gamma_{01}$ and $\log \gamma_{\text{death}}$ for LASSO and GLASSO, respectively

we attached Figure 3.4 to demonstrate how the coefficient estimates shrink for each type of transition as the penalization increases by adjusting the tuning parameter along a one-dimensional scale when using a single LASSO.

Table 3.4 summarizes lists of the selected variables along with their effects on the three transitions based on the three-state models and Cox models. The three-state model generally selects more variables for the onset of cognitive impairment than for transitions to death, no matter which penalties are used. This is because the candidate variables are more closely related to the onset of cognitive impairment rather than risk factors for death. The covariates selected for 0-1 transitions using different penalties largely align with each other. The variables selected by LASSO+GLASSO form a subset of those chosen by LASSO, which are a subset of those selected by two LASSO models. However, for the 1-2 transition process, a single LASSO identifies more variables than the two separate LASSO models. Among these methods, LASSO+GLASSO selects the fewest variables for the 0-

1 transitions and does not select any variables for transitions to death. The minimum SABIC (7664) achieved by the two separate LASSO penalties is slightly lower than that of the single LASSO penalty (7685), which shows the advantage of using two LASSO penalties for greater flexibility, finer tuning, and improved variable selection performance. Notably, LASSO+GLASSO achieves the lowest SABIC (7659) among the three-state models.

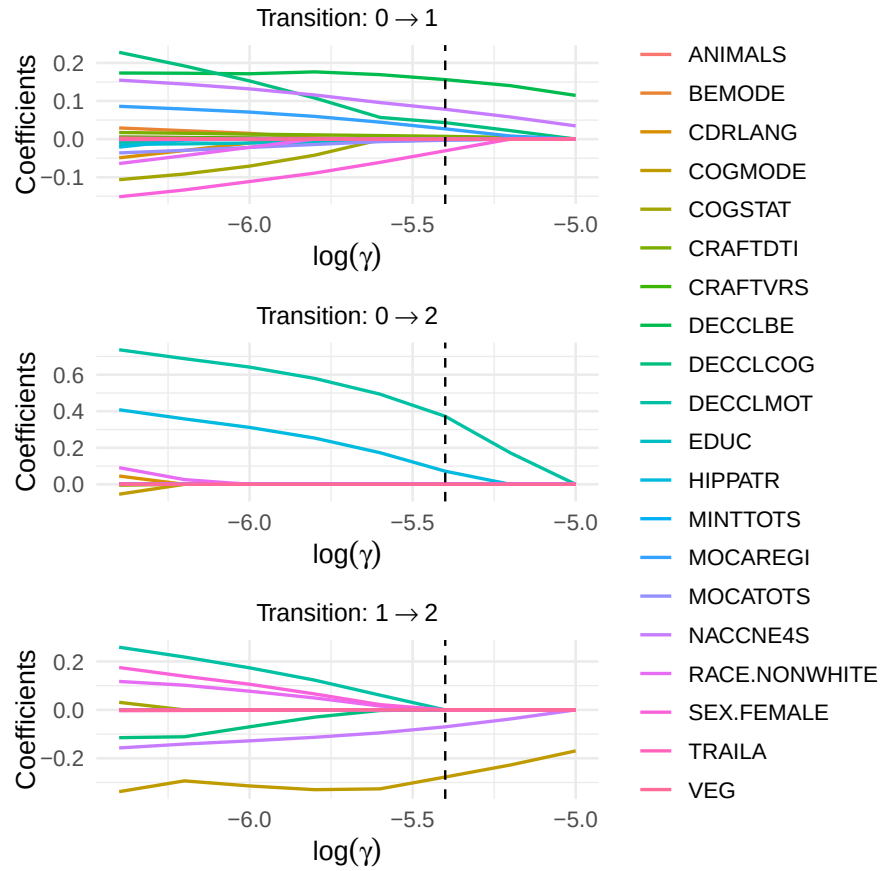


Figure 3.4: Shrinkage estimates of coefficients from penalized EM algorithm with a single LASSO penalty across values of $\log \gamma$

For the transition from the healthy state to cognitive impairment, female sex (SEX.FEMALE)

	LASSO			LASSO ¹			LASSO + GLASSO ²			Cox	
	0 → 1	0 → 2	1 → 2	0 → 1	0 → 2	1 → 2	0 → 1	0 → 2	1 → 2	PFS	OS
SEX.FEMALE	-0.061		0.022	-0.092							
RACE.NONWHITE			0.016							-0.36	
NACCNE4S	0.096		-0.095	0.115		-0.069	0.062				
EDUC	-0.005			-0.008						0.014	
DECCLCOG	0.057			0.111						0.705	0.358
DECCLBE	0.169			0.179			0.152			0.265	
DECCLMOT		0.493	0.063		0.368					0.418	0.355
MOCATOTS	-0.007			-0.016						-0.053	-0.018
COGSTAT	-0.002			-0.049						0.121	
HIPPATR		0.173			0.072					0.367	
CDRLANG				-0.005						0.049	0.247
CRAFTVRS				0.001						-0.027	
VEG										-0.06	-0.046
ANIMALS				0.002						-0.009	-0.007
TRAILA											0.005
CRAFTDTI	0.009			0.011			0.005				
MOCAREGI	0.044			0.059			0.004			-0.043	-0.119
MINTTOTS										0.022	
COGMODE			-0.329			-0.282					
BEMODE											

¹An alternative LASSO with two tuning parameters, each for the $0 \rightarrow 1$ transition and the death process, respectively.

²LASSO for the $0 \rightarrow 1$ process and GLASSO for the $0 \rightarrow 2$ and $1 \rightarrow 2$ process.

Table 3.4: Selected covariates and estimated coefficients by the 3-state model and the Cox model

has a negative effect, indicating that healthy females develop cognitive impairment at a slower rate than males. However, among those who have already developed cognitive impairment, females face a higher risk of death compared to males. Additionally, among the cognitively impaired group, white individuals have a higher risk of death. The presence of APOE e4 alleles (NACCNE4S) is associated with an increased risk of the onset of cognitive impairment, which is consistent with its well-established role as a genetic risk factor for Alzheimer’s disease ([Safieh and others, 2019](#)). Years of education is found to slightly reduce the risk of developing cognitive impairment. Some test scores and clinical judgments have been identified as prognostic indicators of cognitive impairment onset, including one of the neuropsychological battery scores (MOCATOTS), number of abnormal test scores (COGSTAT), extent of clinician-assessed meaningful cognitive impairment (DECCLCOG), extent of behavioral symptoms based on clinical judgment (DECCLBE), and presence of motor symptoms assessed by clinicians (DECCMOT).

3.6 Discussion

In this study, we demonstrated that a penalized EM algorithm can effectively identify important variables within an illness-death model under a mixed observation scheme. Specifically, the entry time to the illness state is subject to interval-censoring, while the time to death is subject to right-censoring. This can result in the entry time to the illness state being interval-censored or, in some cases, the path of a subject being unknown. We proposed an EM algorithm to address the missing information and made a piecewise assumption for the baseline transition intensities. The use of the piecewise constant assumption enables us to transform the conditional expectations in the E-step into a Poisson likelihood and exploit existing R functions for maximization. We have assumed independent delayed entry

throughout - extensions accommodating dependent delayed entry or more formal sample selection schemes are natural and a topic of ongoing work.

Chapter 4

Impact of Model Misspecification with Multistate Processes: Some Results for Regression Analysis and Prediction

4.1 Introduction

4.1.1 Multistate and Alternative Models

Much attention has been given in recent years to the merits of joint modeling of longitudinal markers and failure times ([Ibrahim *and others*, 2010](#)). It has been widely applied to better understand disease progression and associated risks, such as using repeatedly measured hyperglycemia levels as a biomarker for type I diabetes to study the risk of cardiovascular

disease ([Miller and others, 2018](#)), and assessing the association between longitudinal CD4 counts and survival in HIV studies ([Wulfsohn and Tsiatis, 1997](#)). In practice, continuous markers are often discretized using clinically relevant cut-points. This enables jointly modeling of the marker process and the failure process by introducing a discrete state space and allows use of multistate models. See [Cook and others \(2022\)](#) for a description of a joint marker-failure process model of this type. Multistate models are also useful for joint consideration of non-fatal and fatal events. The illness-death process is the simplest such joint model which plays an important role in many clinical contexts, including predicting time to relapse and survival in cancer, identifying prognostic factors for dementia in studies of aging, and among many others ([Touraine and others, 2016](#)).

In joint multistate models fixed covariates can have effects on the failure time but may also affect the transition intensities of the marker process or the non-fatal events. However, multiple incompatible models targeting different marginal features are routinely fitted in practice when interest lies in examining treatment effects on a variety of outcomes, such as absolute risk regression ([Gerds and others, 2012](#)) for time to disease progression, Cox regression ([Cox, 1972](#)) for overall survival, and some alternative models based on composite endpoints. Extensive research has been conducted regarding the relative bias and efficiency of estimators from fitting incompatible marginal models to time-to-event data. [Bernardo and Harrington \(2001\)](#) introduced a framework for sample size calculation under a multiplicative intensity model for recurrent event processes and progressive Markov processes. [Wu and Cook \(2012\)](#) examined the asymptotic and empirical properties of estimators from a Cox regression model for time to first event under multivariate failure processes, and discussed the challenges in interpreting the estimated effects on time to composite endpoints. [Zhong and Cook \(2021\)](#) studied the bias and relative efficiency of time to first event analyses and recurrent event analyses under a broad class of data

generating models and also examined power implications for tests of treatment effects in randomized trials. Direct comparisons between illness-death models and simpler Cox regression estimators have been conducted in several studies. [Leffondré *and others* \(2013\)](#) compared the performance of an illness-death model with a Cox model when the time to the intermediate event is interval-censored, and found that the illness-death process provides a more accurate estimate for the exposure effect on disease progression. [Meira-Machado and Faria \(2014\)](#) compared the finite-sample performance of covariate effect estimates obtained from illness-death models with those from a Cox model incorporating a time-dependent covariate to account for the increased risk of death following entry into the intermediate state.

When the primary goal is prediction, such as predicting time to an event or occurrence of an event by a prespecified time, multistate models offer a flexible and unifying framework. There is a wide range of other approaches available for predicting marginal features of disease processes. When predicting the probability of experiencing an event of interest in the presence of competing risks, a common method is to fit generalized linear models ([Nelder and Wedderburn, 1972](#)) to the cumulative incidence function of the primary event. Estimation for such a model can be based on the Fine-Gray approach ([Fine and Gray, 1999](#)) or direct binomial regression ([Scheike *and others*, 2008](#)). This approach can be applied to general multistate processes other than competing risk scenarios; see [Bühler *and others* \(2024\)](#) for a detailed discussion about the Fine-Gray and direct binomial approaches and the properties of their treatment effect estimators in randomized trials. One can also construct and fit models for state occupancy probabilities at a single time or possibly a sequence of time points using the pseudo-value approach ([Andersen *and others*, 2004](#)).

Our research is motivated by the need to better understand the benefits of risks of fitting multistate models to complex life history processes when interest lies in marginal

features of the process. In particular, Cox proportional hazards models and Fine-Gray competing risk models are often used instead of more comprehensive multistate models such as illness-death processes. We aim to study the impact of model misspecification on biases in regression coefficient estimates and predictive accuracy by both deriving asymptotic properties and evaluating finite-sample performance through simulation studies. The remainder of this chapter is organized as follows. In Section 4.2, we introduce the notation for general multistate models and present the estimating equations for both a Markov illness-death model and a Cox proportional hazards model. Section 4.3 presents the limiting values of regression coefficient estimators derived from Cox regression and illness-death models under various scenarios. In Section 4.4, we examine the impact of model misspecification on predictive accuracy across different prediction goals, including overall survival and disease progression; for the latter, we consider the Fine-Gray model as a comparator to the multistate model. Section 4.5 reports results from simulation studies designed to evaluate the finite-sample performance of estimators derived from Cox regression, Fine-gray models, and multistate analysis in terms of regression coefficient estimation and prediction. In Section 4.6, we compared the three models to predict overall survival and fracture risk in patients with metastatic breast cancer in the control arm. Concluding remarks and directions for future work are provided in Section 4.7.

4.2 Notation

4.2.1 Notation for General Multistate Processes

We first introduce some notation for a general multistate process with state space $\mathcal{S} = \{0, 1, \dots, K + 1\}$. Let $\{Z(s), s \geq 0\}$ denote the process, where $Z(t)$ represents the state

occupied at time t , and the initial state $Z(0) = 0$ with probability one. Let $\mathcal{H}(t) = \{Z(u), 0 \leq u \leq t; X\}$ denote the complete history of the disease process up to time t with a $p \times 1$ fixed covariate vector $X = (X_1, \dots, X_p)'$. We assume transitions from state j to state k are governed by the intensity function

$$\lambda_{jk}(t | \mathcal{H}(t^-)) = \lim_{\Delta t \rightarrow 0} \frac{P(Z(t + \Delta t^-) = k | Z(t^-) = j, \mathcal{H}(t^-))}{\Delta t}, \quad j, k \in \mathcal{S}.$$

Under a Markov model with multiplicative covariate effects, $\lambda_{jk}(t | \mathcal{H}(t^-))$ can be expressed as

$$\lambda_{jk}(t | X) = \lambda_{jk}(t) \exp(X' \beta_{jk}), \quad (4.1)$$

where $\lambda_{jk}(t)$ is the baseline intensity for $j \rightarrow k$ transitions and β_{jk} is the corresponding vector of regression coefficients. Accordingly, we define the cumulative intensity function as $\Lambda_{jk}(t | X) = \Lambda_{jk}(t) \exp(X' \beta_{jk})$ where $\Lambda_{jk}(t) = \int_0^t \lambda_{jk}(s) ds$ is the cumulative baseline intensity function.

We now introduce counting process notation to facilitate discussions of maximum likelihood methods. Consider a sample of n independent processes observed subject to right-censoring, with subscript i indexing individuals. We let C_i represent the right-censoring time, let $Y_i(s) = I(s < C_i)$ indicate whether the disease process is under observation at time s , and $Y_{ij}(s) = I(Z_i(s^-) = j)$ indicate whether individual i is a risk of transitioning out of state j . We further let $\bar{Y}_{ij}(s) = Y_i(s)Y_{ij}(s)$ indicate whether individual i is at risk of making a transition out of state j and under observation for $j \in \mathcal{S}$. Let $dN_{ijk}(s)$ indicate whether a $j \rightarrow k$ transition is made at time s for individual i , and $N_{ijk}(t) = \int_0^t dN_{ijk}(s)$ count the number of $j \rightarrow k$ transitions over $(0, t]$, $i = 1, \dots, n$.

We specify a working modulated Markov model as

$$\gamma_{jk}(t \mid X_i) = \gamma_{jk}(t) \exp(X_i' \psi_{jk}), \quad (4.2)$$

where $\gamma_{jk}(t)$ is the baseline intensity for $j \rightarrow k$ transitions and ψ_{jk} is the corresponding vector of regression coefficients. We refer to this as a “working” model since it may not align with the true data generating process but it may be specified for analysis and model fitting. Using Riemann-Stieltjes integration ([Andersen and others, 2012](#)), the observed data log-likelihood based on the working model is

$$\ell(\theta) = \sum_{i=1}^n \sum_{j \in \mathcal{S}} \sum_{k \in \mathcal{S}} \int_0^\infty \bar{Y}_{ij}(s) [dN_{jk}(s) \log d\Gamma_{ijk}(s) - d\Gamma_{ijk}(s)], \quad (4.3)$$

with $\Gamma_{ijk}(t) = \int_0^t \gamma_{ijk}(s) ds$ and

$$d\Gamma_{ijk}(s) = d\Gamma_{jk}(s) \exp(X_i' \psi_{jk}), \quad (4.4)$$

where $d\Gamma_{jk}(t) = \gamma_{jk}(t) dt$ denotes the baseline increment in the cumulative intensity function, and $\theta = (d\Gamma_{jk}(\cdot), \psi_{jk}', j, k \in \mathcal{S})'$. Taking the partial derivatives of (4.3) with respect to $d\Gamma_{jk}(s)$ and ψ_{jk} gives the score function

$$U_{jk}(s) = \sum_{i=1}^n \bar{Y}_{ij}(s) [dN_{ijk}(s)/d\Gamma_{jk}(s) - \exp(X_i' \psi_{jk})] \quad (4.5)$$

for $d\Gamma_{jk}(s)$ and the score vector

$$U_{jk}(\psi_{jk}) = \sum_{i=1}^n \int_0^\infty \bar{Y}_{ij}(s) [dN_{ijk}(s) - d\Gamma_{jk}(s) \exp(X_i' \psi_{jk})] X_i \quad (4.6)$$

for the regression coefficients ψ_{jk} . Setting (4.5) to zero and solving gives

$$d\tilde{\Gamma}_{jk}(s; \psi_{jk}) = \frac{\sum_{i=1}^n \bar{Y}_{ij}(s) dN_{ijk}(s)}{\sum_{i=1}^n \bar{Y}_{ij}(s) \exp(X_i' \psi_{jk})}. \quad (4.7)$$

Substituting (4.7) into (4.6) gives the partial working score vector for ψ_{jk} (Kalbfleisch and Prentice, 2002)

$$U_{jk}(\psi_{jk}) = \int_0^\infty \bar{Y}_{ij}(s) \left[X_i - \frac{R^{(1)}(s, \psi_{jk})}{R^{(0)}(s, \psi_{jk})} \right] dN_{ijk}(s), \quad (4.8)$$

where $R^{(l)}(s, \psi) = \sum_{i=1}^n \bar{Y}_{ij}(s) X_i^l \exp(X_i' \psi)$ for $l = 0, 1$. By setting (4.8) to zero and solving, we obtain $\hat{\psi}_{jk}$, the estimator for ψ_{jk} . Then by plugging $\hat{\psi}_{jk}$ into (4.7), we obtain the estimator for the baseline parameters $d\Gamma_{jk}(t)$ given by

$$d\hat{\Gamma}_{jk}(s) = \frac{\sum_{i=1}^n \bar{Y}_{ij}(s) dN_{ijk}(s)}{\sum_{i=1}^n \bar{Y}_{ij}(s) \exp(X_i' \hat{\psi}_{jk})}, \quad s > 0, j, k \in \mathcal{S}. \quad (4.9)$$

The maximum partial likelihood estimator $\hat{\psi}_{jk}$ is a consistent estimator of ψ_{jk}^* (Struthers and Kalbfleisch, 1986) which solves

$$\int_0^\infty \left[r_{jk}^{(1)}(s) - \frac{r_{jk}^{(1)}(s; \psi_{jk})}{r_{jk}^{(0)}(s; \psi_{jk})} r_{jk}^{(0)}(s) \right] = 0 \quad (4.10)$$

where $r_{jk}^{(l)}(s) = \mathbb{E}[\bar{Y}_{ij}(s) X_i^l dN_{ijk}(s)]$ and $r_{jk}^{(l)}(s; \psi_{jk}) = \mathbb{E}[\bar{Y}_{ij}(s) X_i^l \exp(X_i' \psi_{jk})]$ for $l = 0, 1$, and the expectation is taken with respect to the true joint distribution of the event process, covariate process, and observation process. The limiting value of $d\hat{\Gamma}_{jk}(s)$ is obtained by setting the expectation of the score function (4.5) to zero with $\hat{\psi}_{jk}$ replaced by its limiting

value ψ_{jk}^* and is given by

$$d\Gamma_{jk}^*(s) = \frac{\mathbb{E} [\bar{Y}_{ij}(s) dN_{ijk}(s)]}{\mathbb{E} [\bar{Y}_{ij}(s) \exp(X_i' \psi_{jk}^*)]} = \frac{r_{jk}^{(0)}(s)}{r^{(0)}(s; \psi_{jk}^*)}. \quad (4.11)$$

See [Struthers and Kalbfleisch \(1986\)](#) and [Lin and Wei \(1989\)](#) for the derivation of the robust large sample variance matrix for $\hat{\psi}_{jk}$ and how to estimate it.

4.2.2 An Illness-death Process and Two Models

Here we consider an illness-death process in which state 0 represents a healthy state, state 1 denotes an intermediate state, usually indicating progression of disease, and state 2 is an absorbing state typically representing death. The model allows for transitions from the healthy state to the death state, either directly or via the intermediate illness state. Figure 4.1 provides two different representations of the illness-death process. The first representation is a standard multistate diagram involving the healthy, progression, and death states. The second representation distinguishes between post-progression death and progression-free death, which is often useful to do; see the discussion in Section 4.4. In what follows, we discuss two commonly-used models for analyzing data from the illness-death process, including a compatible three-state model and the Cox proportional hazards model.

If we adopt a compatible three-state Markov model with multiplicative effects of X , all the estimating functions and the forms of estimators shown in Section 4.2.1 apply here for $j < k$, $j, k = 0, 1, 2$. When interest lies in the overall survival time T_2 , defined as the time of entry into state 2 in Figure 4.1(a), it is common to use a Cox proportional hazards model. The model assumes that the failure process is governed by the hazard function

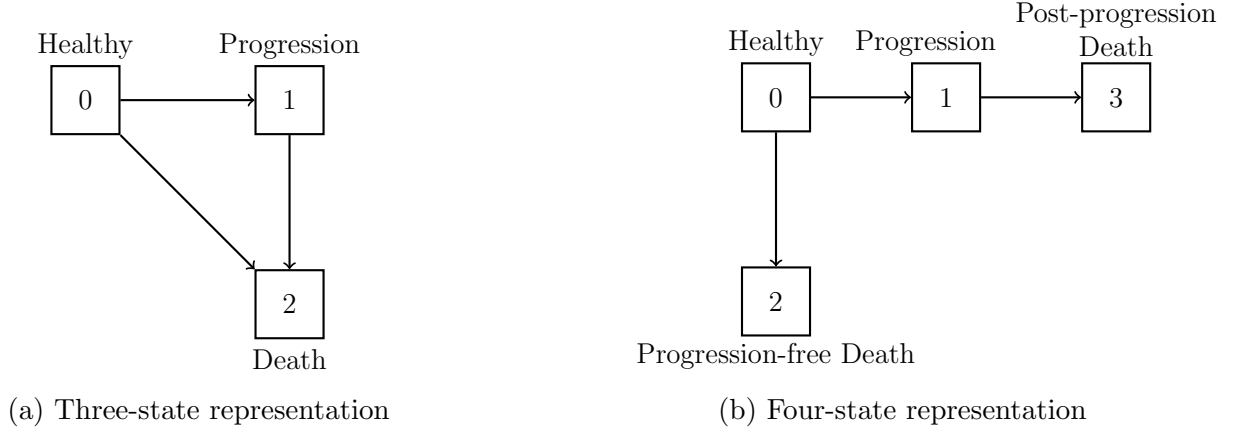


Figure 4.1: Two representations of the illness-death process.

with multiplicative covariate effects, which takes the form

$$\gamma(t | X) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T_2 < t + \Delta t | T_2 \geq t, X)}{\Delta t} = \gamma_0(t) \exp(X' \psi),$$

where $\gamma_0(t)$ denotes the baseline hazard function, and ψ is the vector of regression coefficients representing the effect of covariates X on the hazard of overall survival. The Cox regression essentially corresponds to a two-state model, where states 0 and 1 in Figure 4.1(a) are collapsed into one “alive” state. We let $dN_2(s)$ denote that state 2 is entered at time s . Define at-risk indicators as follows: Let $Y^\dagger(s) = I(Z(s^-) < 2)$ indicate that the individual remains at risk of death at time s , and $\bar{Y}(s) = Y(s)Y^\dagger(s)$ as the indicator that an individual is at risk of death and under observation at time s . The estimating equation for the limiting value ψ^* is given by (Struthers and Kalbfleisch, 1986)

$$\int_0^\infty \left[r^{(1)}(s) - \frac{r^{(1)}(s; \psi)}{r^{(0)}(s; \psi)} r^{(0)}(s) \right] = 0, \quad (4.12)$$

where $r^{(l)}(s) = \mathbb{E} [\bar{Y}(s) X^l dN_2(s)]$, and $r^{(l)}(s; \psi) = \mathbb{E} [\bar{Y}(s) X^l \exp(X' \psi)]$ for $l = 0, 1$, and

the limiting value of the estimator for the baseline hazard $d\Gamma_0(s)$ is

$$d\Gamma_0^*(s) = \frac{\mathbb{E} [\bar{Y}(s) dN_2(s)]}{\mathbb{E} [\bar{Y}(s) \exp(X'\psi^*)]} = \frac{r^{(0)}(s)}{r^{(0)}(s, \psi^*)}. \quad (4.13)$$

4.3 Limiting Values for Estimators of Regression Coefficients

Here we examine the limiting values of regression coefficients under various scenarios, focusing on two main scenarios. First, we study the limiting values of estimators for Cox regression models when the true data generating process is three-state Markov illness-death model. Second, we explore potential risks associated with using a three-state Markov model when it is based on invalid assumptions such as a) omitting important covariates and b) assume a Markov model when it is semi-Markov.

4.3.1 Limiting Values of Cox Regression Estimators

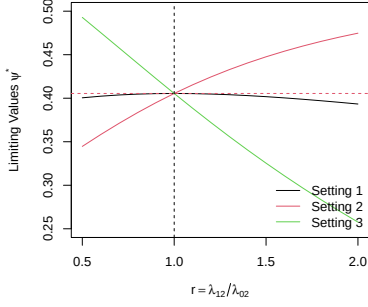
Here we assume that the true process is a time-homogeneous Markov illness-death model such that $\lambda_{jk}(t) = \lambda_{jk}$ for $j < k, j, k = 0, 1, 2$, and consider some representative settings to evaluate the limiting values of estimators for Cox regression coefficients. We start with the simplest case involving a single binary covariate and investigate six settings shown in Table 4.1. In Scenario A, the effect of X_1 on the mortality rates are the same - having $X_1 = 1$ increases the $0 \rightarrow 2$ and $1 \rightarrow 2$ transition intensities by a factor of 1.5. In Setting A1, there is no effect of X_1 on the $0 \rightarrow 1$ transition intensity. In Setting A2, individuals with $X_1 = 1$ has 1.5 times higher intensity for $0 \rightarrow 1$ transitions compared to those with $X_1 = 0$, and in Setting A3, those with $X_1 = 1$ have 50% lower intensity for $0 \rightarrow 1$ transitions. In

scenario B the effect of X_1 is different for the $0 \rightarrow 2$ and $1 \rightarrow 2$ intensities. In Setting B1, the effects of X_1 is different on the $0 \rightarrow 2$ and $1 \rightarrow 2$ transition intensities are in the opposite direction - the covariate has a negative effect on the $0 \rightarrow 2$ transition intensity ($\log 0.5$) and a positive effect on the $1 \rightarrow 2$ transition intensity ($\log 2$). The effect on the $0 \rightarrow 1$ transition intensity is positive at $\log 2$. In Setting B2, X_1 has no effect on the $0 \rightarrow 2$ transition intensity, whereas it reduces the intensity of the $0 \rightarrow 1$ transition by half, and doubles the intensity of the transition from state 1 to state 2. In Setting B3, the covariate has a negative effect on the $0 \rightarrow 1$ and $1 \rightarrow 2$ transition intensity ($\log 0.5$) and a positive effect on the $0 \rightarrow 2$ transition intensity ($\log 2$).

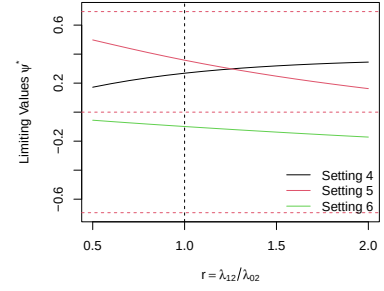
We assume the baseline intensities for transitions to death are proportional in all settings, such that $\lambda_{12} = r\lambda_{02}$, where r is a constant. When $r > 1$, there is an elevated baseline risk of death when an individual is in the intermediate state 1 compared to the initial state. We let T_{01} and T_{02} be (possible latent) $0 \rightarrow 1$ and $0 \rightarrow 2$ transition times. For specified r the baseline intensities λ_{01} and λ_{02} can be found which satisfy $E_X[P(T_{01} < T_{02} | X)] = 0.7$ and $E_X[P(T_2 < C | X)] = 0.8$, which correspond to a 70% probability of making a direct $0 \rightarrow 1$ transition and a net right censoring rate of 20% for the entry time to state 2. For simplicity, we ignore random censoring and assume a fixed administrative censoring time of $\tau = 1$.

Transition	Parameter	SCENARIO A			SCENARIO B		
		A1	A2	A3	B1	B2	B3
Progression	β_{01}	0	$\log 1.5$	$\log 0.5$	$\log 2$	$\log 0.5$	$\log 0.5$
Progression-free death	β_{02}	$\log 1.5$	$\log 1.5$	$\log 1.5$	$\log 0.5$	0	$\log 2$
Post-progression death	β_{12}	$\log 1.5$	$\log 1.5$	$\log 1.5$	$\log 2$	$\log 2$	$\log 0.5$

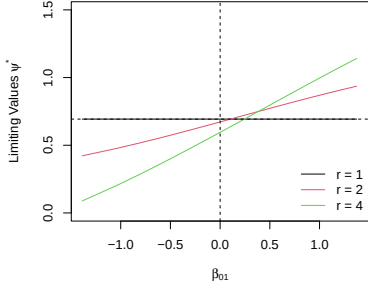
Table 4.1: Coefficients for five settings involving a single binary covariate $X_1 \sim \text{Bernoulli}(0.5)$ with constraints $\lambda_{12}/\lambda_{02} = r$



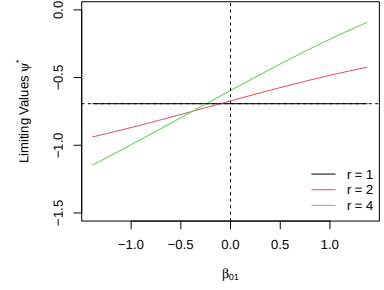
(a) The red dashed line is $\log 1.5$, representing the conceptualized true effect on the risk of death in Scenario A.



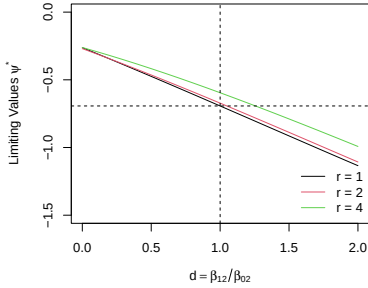
(b) The red dashed lines represent true effects on transitions in Scenario B.



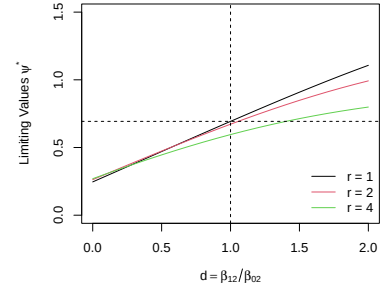
(c) The black dashed line is $\log 2$, representing the true effect in Setting 7 ($\beta_{02} = \beta_{12} = \log 2$).



(d) The black dashed line is $\log 0.5$, representing the true effect in Setting 8 ($\beta_{02} = \beta_{12} = \log 0.5$).



(e) The black dashed line is $\log 0.5$, representing the true effect on $0 \rightarrow 2$ transition in Setting 9 ($\beta_{01} = 0, \beta_{02} = \log 0.5$).



(f) The black dashed line is $\log(2)$, representing the true effect on $0 \rightarrow 2$ transition in Setting 10 ($\beta_{01} = 0, \beta_{02} = \log 2$).

Figure 4.2: Limiting values for the regression coefficient estimator under Cox regression involving a binary covariate $X_1 \sim \text{Bernoulli}(0.5)$. Panels (a)-(b): Settings A and B, varying $r = \lambda_{12}/\lambda_{02}$; (c)-(d): Settings 7-8, fix $r = \lambda_{12}/\lambda_{02} = 1, 2$ or 4 , varying over β_{01} ; (e)-(f): Settings 9-10, varying $d = \beta_{12}/\beta_{02}$.

Figure 4.2(a) shows the limiting value of the estimator for ψ from fitting a Cox model varies as a function of $r = \lambda_{12}/\lambda_{02}$ in Scenario A for Setting A1-A3. In Figure 4.2(a), we observe that the three curves corresponding to Settings A1-A3 intersect at $r = 1$ where the limiting value equals the “true” effect because here the risk of death is the same for transitions from the healthy or intermediate state. When $r \neq 1$, if we conceptualize $\log 1.5$ as the “true” effect on the risk of death, the Cox model yields attenuated estimates in Setting A1. In Setting A2, it underestimates the effect when $r < 1$ and overestimates it when $r > 1$, whereas in Setting A3 we observe the opposite pattern - it overestimates the effect when $r < 1$ and underestimates it when $r > 1$. For Scenario A, the bias is 0 when $r = 1$ and increases as r departs from 1. For Setting A1 where the effect on $0 \rightarrow 1$ transition intensity is 0, the bias remains small as r deviates from 1. However, in Settings 2 and 3, Cox model estimators can result in substantially biased estimates of the effect on the death intensities.

In Scenario C, X_1 has a common effect on both the $0 \rightarrow 2$ and $1 \rightarrow 2$ transition intensities, which are positive ($\log 2$) in Setting C1 or negative ($\log 0.5$) in Setting C2, respectively. Figure 4.2(c)-(d) shows the limiting value for regression coefficient estimator as a function of β_{01} values when $\beta_{02} = \beta_{12}$ and $r = 1$. In both settings, the limiting value ψ^* remains equal to the true common effect β_{02} (or β_{12}) across values of β_{01} ranging from $\log 0.25$ to $\log 4$ when $r = 1$, which aligns with our findings in Figure 4.2(a). This arises because under the time-homogeneous Markov illness-death model, the proportional hazards assumption holds, so the Cox model is not misspecified. To see this, we show that under the assumed time-homogeneous illness-death process, $\mathcal{F}(t | X) = P(Z(t) < 2 | X)$

is given by

$$\begin{aligned}\mathcal{F}(t | X) &= \exp [-\Lambda_{01}(t | X) - \Lambda_{02}(t | X)] \\ &+ \int_0^t \exp \{ - [\Lambda_{01}(s | X) + \Lambda_{02}(s | X)] \} \lambda_{01}(s | X) \exp (-\Lambda_{12}(s, t | X)) ds.\end{aligned}\tag{4.14}$$

If $\beta_{02} = \beta_{12} = \beta$ and $\lambda_{02} = \lambda_{12} = \lambda$, this can be simplified to $\exp(-\lambda \exp(X'\beta)t)$. The baseline survival function takes the form $\mathcal{F}(t | X = 0) = \exp(-\lambda t)$. Since $\mathcal{F}(t | X) = \mathcal{F}(t | X = 0)^{\exp(X'\beta)}$, the proportional hazards assumption holds. However, as illustrated in Figure 4.2(c)-(d), when $r \neq 1$, Cox regression yields inconsistent estimates with the limiting value of the regression coefficient estimator increasing as β_{01} increases.

Figure 4.2(b) shows the asymptotic value of the estimator for ψ in Setting B1-B3. We include three red reference lines to indicate the true effects on the $0 \rightarrow 2$ and $1 \rightarrow 2$ transition intensities specified in each setting. In Setting B1, the limiting value increases with r , while in Setting B2 and Setting B3 they decrease. In all three settings, the estimated effect from Cox regression lies between the corresponding “true” effects on the $0 \rightarrow 2$ and $1 \rightarrow 2$ transitions when r is between 0.5 and 2.

In Scenario D we express the covariate effect on the $1 \rightarrow 2$ transition as a multiple of its effect on the $0 \rightarrow 2$ transition such that $\beta_{12} = d\beta_{02}$. We assume the covariate has no effect on $0 \rightarrow 1$ transitions ($\beta_{01} = 0$) and consider $r = 1, 2$, and 4. Figure 4.2(e)-(f) present the limiting values of the Cox regression coefficient estimator as a function of d . In Setting D1, the effect on the $0 \rightarrow 2$ transition is negative ($\beta_{02} = \log 0.5$), shown in Figure 4.2(e), and in Setting D2, the effect is positive ($\beta_{02} = \log 2$), shown in Figure 4.2(f). We observe that the trend in the limiting values of the estimator with respect to d depends on the direction of the effect on the $0 \rightarrow 2$ transition. Specifically, when $\beta_{02} < 0$, the limiting value decreases

as d increases, and the opposite trend is observed when $\beta_{02} > 0$. This pattern is observed for $r = 1, 2$, and 4. In addition, for a fixed β_{02} , a larger r generally corresponds to higher limiting values when $\beta_{02} < 0$ and lower limiting values when $\beta_{02} > 0$ in the range $(0, 2]$ for d .

In summary, Cox regression is generally misspecified when applied to data generated from an illness-death process, except under some specific conditions: when the underlying illness-death process is Markov, the baseline intensities for transitioning to state 2 from both states 0 and 1 are equal, and the covariate effects on the $0 \rightarrow 2$ and $1 \rightarrow 2$ transitions are identical. As long as these three criteria are satisfied, the limiting value of the regression coefficient estimated from the Cox model equals the common covariate effect on the $0 \rightarrow 2$ and $1 \rightarrow 2$ transitions and is not influenced by the effect on the $0 \rightarrow 1$ transition. In all other settings we investigated, the Cox model estimate depends on the value of r and the covariate effects on all three transitions.

4.3.2 Misspecified Multistate and Cox Models

Here we examine the implications of assuming a Markov illness-death model when it is misspecified with a focus on the limiting values of regression coefficient estimators; we also assess the limiting values of Cox regression coefficient estimators under similar data generation models. Focus is primarily on scenarios where the Markov assumption is violated because

- 1) Important covariates are omitted, and
- 2) the underlying process features a semi-Markov $1 \rightarrow 2$ intensity

Implications of omitting a covariate

First, we discuss the implications of omitting a covariate when fitting a Markov illness-death model or a Cox model. For simplicity, we assume there are two binary covariates X_1 and X_2 with $p_{ij} = P(X_1 = i, X_2 = j)$. We assume $P(X_j = 1) = 0.5$ marginally and consider their association as characterized by an odds ratio (OR) with $OR = p_{11}p_{00}/(p_{10}p_{01})$ set to $OR = 0.5, 1$, or 2 corresponding to negative association, independence, and positive association, respectively. To focus on the effect of covariate omission, we consider a special setting in which both X_1 and X_2 have common effects across all three transitions and $r = \lambda_{12}/\lambda_{02} = 1$. We set the effect of X_1 to $\log 1.5$, and vary the effect of X_2 to examine how the limiting values change as a function of the effect of X_2 . Derivations for the limiting values are provided in Appendix C.2 for the regression coefficients in the Markov illness-death model

Figure 4.3 shows the limiting values of the regression coefficients varying as a function of the common effect of the omitted covariate X_2 expressed as the regression coefficient. We report ψ_{01}^* , ψ_{02}^* , and ψ_{12}^* , the limiting values for the regression coefficient estimators of the three transition rates under a working Markov illness-death model given only X_1 , along with ψ^* , the limiting value of the Cox regression coefficient estimator. Figure 4.3 contains three panels, which correspond to the dependence between the covariates set as $OR = 0.25, 1$, and 2 , respectively. We observe that the implications of omitting X_2 vary by both the common effect of X_2 and the association between X_1 and X_2 . If X_1 is independent of X_2 , omitting X_2 consistently leads to attenuation in the limiting value of the X_1 effect estimator for both the Cox model and the Markov illness-death model across three transitions. When X_1 and X_2 are positively associated ($OR = 2$), omitting X_2 results in a larger limiting value for the estimator of the effect of X_1 when X_2 has a positive effect,

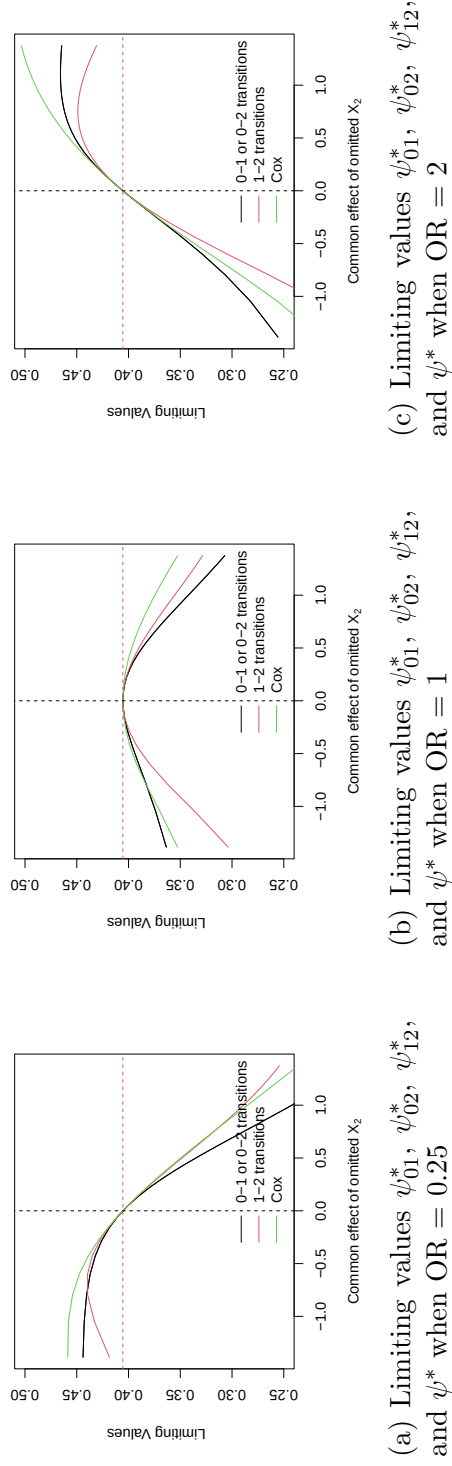


Figure 4.3: Limiting values of the estimators for regression coefficients β_{01} , β_{02} , and β_{12} under the Markov illness-death model and of ψ under Cox regression plotted as functions of the common regression coefficient for X_2 on all three transitions when X_2 is omitted from the model. Note X_1 has common regression coefficient of $\log 1.5$ across all three transitions and $r = \lambda_{12}/\lambda_{02} = 1$. In each panel, the horizontal red dashed line indicates the true effect of X_1 , while the black, red, and green curves correspond to limiting values ψ_{01}^* (or ψ_{02}^*), ψ_{12}^* , and ψ^* , respectively.

and a smaller limiting value when X_2 has a negative effect. The trend is reversed when X_1 and X_2 are negatively associated; see Figure 4.3(a).

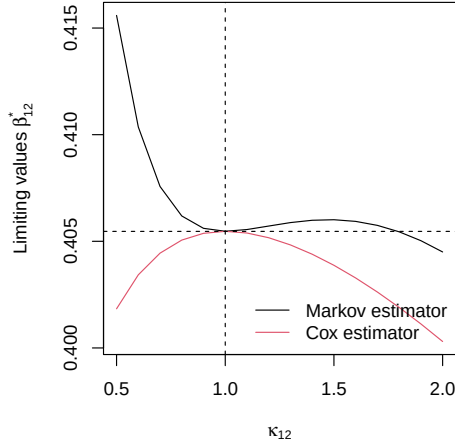
Implications of a Semi-Markov $1 \rightarrow 2$ Intensity

Now we consider the scenario where the true process follows a semi-Markov illness-death model while a Markov model or a Cox model for overall survival time is fitted. To isolate the impact of each type of misspecification, we consider a data generating process involving a single covariate X_1 which is Bernoulli(0.5) - both the Markov illness-death model and the Cox regression include X_1 (i.e., no important covariates are omitted). We let W denote the sojourn time in state 1 and assume it has a survival function

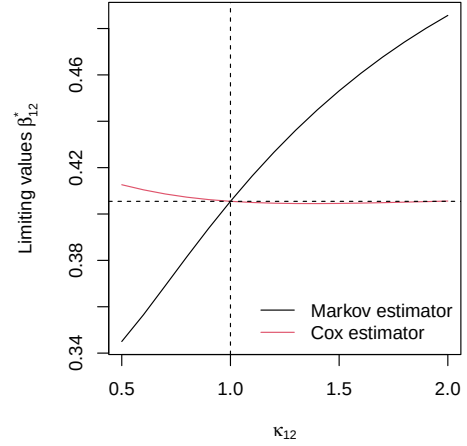
$$\mathcal{F}(w \mid X_1) = \exp [-(\lambda_{12}w)^{\kappa_{12}}] \exp(X_1'\beta_{12}),$$

where λ_{12} and κ_{12} are the rate and shape parameters, respectively. A positive κ_{12} corresponds to an increased risk of making a transition out of state 1 to state 2 over time while negative values correspond to a decreasing risk as expressed by the hazard function, and when $\kappa_{12} = 1$, a Markov illness-death process is retrieved. We examine how the limiting value of the regression coefficient estimator for $1 \rightarrow 2$ transitions from the working Markov illness-death process and the Cox model as a function of κ_{12} .

We consider two settings here. In both Settings 11 and 12 X_1 has a common effect on the two transitions to death so $\beta_{02} = \beta_{12}$. When X_1 has no effect on the $0 \rightarrow 1$ transitions, both models yield estimators with large asymptotic biases when $\kappa_{12} \neq 1$, but depend on κ_{12} in different ways. The Cox regression consistently yields attenuated effects, whereas the limiting value of the estimator for β_{12} in the Markov illness-death model decreases as κ_{12} increases. When X_1 has a positive effect of $\log 1.5$ on the $0 \rightarrow 1$ transitions, estimator



(a) The horizontal black dashed line is $\log 1.5$, representing the true effects on $1 \rightarrow 2$ transition intensities; $\beta_{01} = 0, \beta_{02} = \beta_{12} = \log 1.5$ in Setting 11.



(b) The horizontal black dashed lines are $\log 1.5$ representing the true effects on $1 \rightarrow 2$ transition intensities; $\beta_{01} = \beta_{02} = \beta_{12} = \log 1.5$ in Setting 12.

Figure 4.4: Limiting values of regression coefficient estimator of β_{12} under Markov illness-death model and ψ under a Cox regression involving a single binary covariate $X_1 \sim \text{Bernoulli}(0.5)$, varying over κ_{12} in Settings 11 and 12.

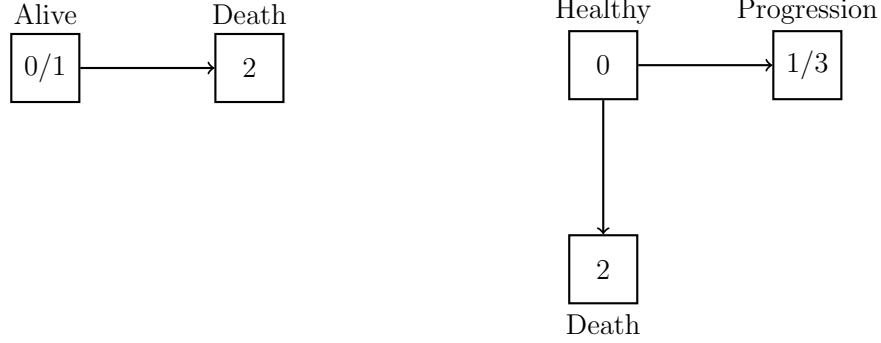
from the working Markov model leads to substantial asymptotic biases in the regression coefficient estimator for the $1 \rightarrow 2$ transitions when κ_{12} is far from 1, whereas the Cox model estimator is close to the conceptualized true value of effects on the risk of death given here by $\beta_{02} = \beta_{12} = \log 1.5$.

4.4 Prediction Performance under Misspecified Multistate Models

Multistate models are also useful for prediction and here we consider the possible utility of them when they may be misspecified. Assessing predictive accuracy plays a central role in the developing, evaluation, and comparison of predictive models. In the context of time-to-event analysis, there are two types of predictors: point predictors and probabilistic predictors. A point predictor gives a single predicted value of the response, while a probabilistic predictor gives a probability of the response. Measures reflecting how close predictors are to actual outcomes (i.e. overall performance) address two properties: discrimination and calibration. Discrimination refers to how well the model differentiates those at higher risk of having an event from those at lower risk. Calibration refers to the agreement between observed outcomes and absolute risk estimates. Standard prediction metrics for binary or continuous outcomes have been adapted to handle censored data using methods such as inverse probability of censoring weighting (IPCW) and imputation ([Schemper and Henderson, 2000](#); [Uno *and others*, 2011](#); [Gerds and Schumacher, 2006](#); [Lawless and Yuan, 2010](#)).

Here we discuss different prediction goals and evaluate the predictive performance of working models using large sample theory and simulation studies. The primary data generating process is a Markov model and we consider this as one of the models on which to base predictive inference. We also consider simple models (e.g., Cox models or generalized transformation models) that may be adapted to estimate particular marginal features. To evaluate predictive accuracy, we use the Brier score ([Glenn *and others*, 1950](#)), which measures the distance between the probabilistic prediction and the actual outcome. The Brier score can take different forms depending on the prediction objective in time-to-event anal-

ysis. We consider two primary predictive goals and two modeling strategies aligned with these two goals: (1) Predicting overall survival, and (2) Predicting disease progression.



(a) A simplified two-state model for predicting overall survival (b) A simplified competing risk process for predicting disease progression

Figure 4.5: Simplified multistate diagrams considered for the purpose of predicting specific marginal features

Predicting overall survival via Cox regression is natural modeling despite the fact it ignores the intermediate state, see Figure 4.5(a). To predict disease progression we consider Figure 4.5(b) and the Fine-Gray model (Fine and Gray, 1999), which treats progression as the primary event of interest and treats progression-free death as a competing event. Both prediction goals can also be addressed through use of a unified multistate model. We therefore aim to compare the predictive performance of the misspecified Cox model and the misspecified Fine-Gray model against that of a Markov illness-death model, acknowledging that there could be settings we do not consider in which the Cox model or the Fine-Gray model is correct but the 3-state model formulated for prediction is not. We therefore also assess the robustness of all three models under scenarios where the 3-state model is misspecified.

4.4.1 Prediction of Overall Survival

To explore the relative predictive performance of the various methods, we consider two simple settings involving either a single binary covariate or a 5×1 vector of independent binary covariates. Covariates with a finite support ensure fast calculations of limiting values, but the methods described in what follows can handle more complex predictors in theory. Let $\mathcal{F}(t | X) = P(Z(t) < 2 | X)$ denote the true survival function of T_2 given covariate X , and $\widehat{\mathcal{F}}_n(t | x)$ denote a general estimator of the survival function based on a sample of n independent observations. The mean squared error of the estimator is defined as ([Gerds and Schumacher, 2006](#))

$$MSE(t, \widehat{\mathcal{F}}_n, \mathcal{F}) = E_X \left\{ E_{T_2} \left[\widehat{\mathcal{F}}_n(t | X) - I(T_2 > t | X) \right]^2 | X \right\}, \quad (4.15)$$

and the expected mean squared error, or the expected Brier score, is defined as the limiting value of the $MSE(t, \widehat{\mathcal{F}}_n, \mathcal{F})$, which is given by

$$\begin{aligned} BS(t) &= E_X \left\{ E_{T_2} [\mathcal{F}^*(t | X) - I(T_2 > t | X)]^2 | X \right\} \\ &= \mathbb{E} [\mathcal{F}^*(t | X) - I(T_2 > t | X)]^2 \end{aligned} \quad (4.16)$$

where $\mathcal{F}^*(t | X)$ is the limit of $\widehat{\mathcal{F}}_n(t | X)$ such that

$$\sup_{t>0} \left| \int [\mathcal{F}^*(t | x) - \widehat{\mathcal{F}}_n(t | x)] dF_X(x) \right| \xrightarrow{\text{a.s.}} 0$$

and $F_X(x)$ is the cumulative distribution function of X ([Schoop and others, 2011](#)). Note that (4.16) can be decomposed as

$$BS(t) = \mathbb{E} [\mathcal{F}^*(t | X) - \mathcal{F}(t | X)]^2 + \mathbb{E} [\mathcal{F}(t | X)(1 - \mathcal{F}(t | X))],$$

where the first term, referred to as imprecision (Hand, 1997), quantifies the distance between the limit of estimated survival function and the true survival function. The second term is called inseparability (Hand, 1997), and represents the irreducible part of the Brier score that remains even when $\hat{\mathcal{F}}_n(t | X)$ is obtained based on assumptions consistent with the true survival function $\mathcal{F}(t | X)$.

In what follows, we show how the limiting value of the survival function estimated based on a Cox model deviates from the true survival function for each value of the binary covariate and how this discrepancy leads to differences in the expected Brier score over time. Specifically, we compare the expected Brier score based on the limiting value of the estimated survival function from a fitted Cox regression model given a single binary covariate against the true survival function. The estimated survival function from a fitted Cox model given a general fixed covariate vector X is denoted as

$$\hat{\mathcal{F}}_{\text{Cox}}(t | X) = \exp \left[- \int_0^t d\hat{\Gamma}_0(s) \exp(X' \hat{\psi}) \right], \quad (4.17)$$

where $d\hat{\Gamma}_0(s)$ is the estimated baseline hazard at time s in the Cox regression, and $\hat{\psi}$ is the estimated regression coefficient. The limiting value for the estimated survival function is given by

$$\mathcal{F}_{\text{Cox}}^*(t | X) = \exp \left[- \int_0^t d\Gamma_0^*(s) \exp(X' \psi^*) \right], \quad (4.18)$$

where ψ^* is the limiting value for the estimated regression coefficient $\hat{\psi}$ under the Cox model, and $d\Gamma_0^*(t)$ is the limiting value of $d\hat{\Gamma}_0(s)$ given by (4.13).

We adopt the 4-state representation of the illness-death process as shown in Figure 4.1(b) for the following discussion. For a fitted illness-death model, the estimated

survival function is given by

$$\widehat{\mathcal{F}}_{\text{AJ}}(t \mid X) = \widehat{P}(Z(t) < 2 \mid X) = \widehat{P}(Z(t) = 0 \mid X) + \widehat{P}(Z(t) = 1 \mid X), \quad (4.19)$$

where the probabilities $\widehat{P}(Z(t) = 0 \mid X)$ and $\widehat{P}(Z(t) = 1 \mid X)$ are estimated with the Aalen–Johansen estimator ([Aalen and Johansen, 1978](#)) as follows. Let $\mathbb{P}(0, t \mid X)$ denote the conditional transition probability matrix containing (k, l) entry $P(Z(t) = l \mid Z(0) = k, X)$ for $(k, l) \in \{(0, 1), (0, 2), (1, 3)\}$. Let $\mathbb{Q}(s \mid X)$ denote the matrix of cumulative non-homogeneous transition rates for the four-state Markov process in Figure 4.1(b). We consider time-homogeneous transition rates for the illness-death process and let $\mathbb{Q}(X)$ be the matrix of the transition intensities; The true transition probability matrix is computable as $\mathbb{P}(0, t \mid X) = \exp(\mathbb{Q}(X)t)$ ([Cox, 2017](#)), and the `expm` function in *R* can be used for computation.

Given estimates of the baseline transition rates $d\Gamma_{jk}(t)$ and regression coefficients ψ_{jk} , let $d\widehat{\Gamma}_{jk}(t \mid X) = d\widehat{\Gamma}_{jk}(t) \exp(X'\widehat{\psi}_{jk})$ for transitions for $(j, k) \in \{(0, 1), (0, 2), (1, 3)\}$, as described in Section 4.2.1. Then let

$$d\widehat{\mathbb{Q}}(t \mid X) = \begin{pmatrix} -d\widehat{\Gamma}_{01}(t \mid X) - d\widehat{\Gamma}_{02}(t \mid X) & d\widehat{\Gamma}_{01}(t \mid X) & d\widehat{\Gamma}_{02}(t \mid X) & 0 \\ 0 & -d\widehat{\Gamma}_{12}(t \mid X) & 0 & d\widehat{\Gamma}_{12}(t \mid X) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \quad (4.20)$$

The estimated transition probability matrix can be obtained through product integration

(Cook and Lawless, 2018) as

$$\widehat{\mathbb{P}}(0, t \mid X) = \prod_{(0, t]} [\mathbb{I} + d\widehat{\mathbb{Q}}(s \mid X)] . \quad (4.21)$$

With $P(Z(0) = 0 \mid X) = 1$, the estimated $(K + 1) \times (K + 1)$ transition probabilities $\widehat{P}(Z(t) = k \mid Z(0) = 0, X)$ corresponding to the $(1, k + 1)$ entry of $\widehat{\mathbb{P}}(0, t \mid X)$, $k = 0, \dots, K$ can be used to estimate, and based on these we can evaluate $\widehat{\mathcal{F}}_{\text{AJ}}(t \mid X)$ given by (4.19). The limiting value for it is given by

$$\mathcal{F}_{\text{AJ}}^*(t \mid X) = P^*(Z(t) < 2 \mid X) = P^*(Z(t) = 0 \mid X) + P^*(Z(t) = 1 \mid X), \quad (4.22)$$

where $P^*(Z(t) = 0 \mid X)$ and $P^*(Z(t) = 1 \mid X)$ can be obtained from the limiting value of the transition probability matrix,

$$\mathbb{P}^*(0, t \mid X) = \prod_{(0, t]} [\mathbb{I} + d\mathbb{Q}^*(s \mid X)] ,$$

where $d\mathbb{Q}^*(s \mid X)$ is the limiting value of $d\widehat{\mathbb{Q}}(s \mid X)$ obtained by replacing $d\Gamma_{jk}(t \mid X)$ with $d\Gamma_{jk}^*(t \mid X)$ in (4.20), and $d\Gamma_{jk}^*(t \mid X) = d\Gamma_{jk}^*(t) \exp(X' \psi_{jk}^*)$, $(j, k) \in \{(0, 1), (0, 2), (1, 2)\}$. Alternatively, these conditional probabilities can be expressed explicitly as:

$$P^*(Z(t) = 0 \mid X) = \exp \left\{ - \int_0^t [d\Gamma_{01}^*(u \mid X) + d\Gamma_{02}^*(u \mid X)] \right\}$$

and

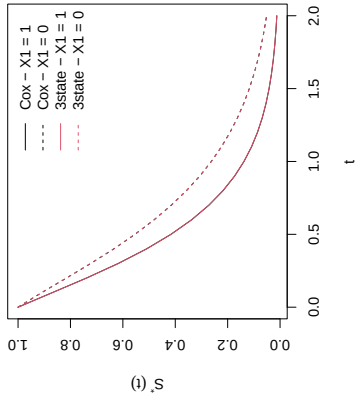
$$P^*(Z(t) = 1 \mid x) = \int_0^t \exp \left\{ - \int_0^s [d\Gamma_{01}^*(u \mid X) + d\Gamma_{02}^*(u \mid X)] \right\} d\Gamma_{01}^*(s \mid X) \exp \left[- \int_s^t d\Gamma_{12}^*(u \mid X) \right] .$$

When the 3-state model is correctly specified, the limiting survival function reduces to the true survival function $\mathcal{F}(t | X) = P(Z(t) < 2 | X)$.

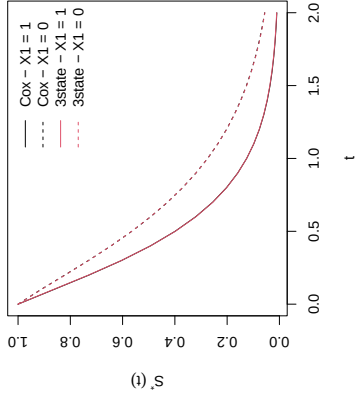
Prediction with A Single Binary Covariate

Here we consider the case with a single binary covariate. We consider Settings 1-6 in Table 4.1 with $r = \lambda_{12}/\lambda_{02} = 2$. The comparison is done regarding the limiting value of estimated survival functions for each value of X_1 , and the expected Brier score is evaluated at the limiting value of the estimated survival function, denoted as $BS(t)$, for the Cox regression and the Markov illness-death model. The results are shown in Figures 6 and 7, respectively. In Figure 4.6, we observe only minor differences between the limiting values for the estimated survival functions based on the Cox model and the true Markov illness-death model when $r = 2$ - a setting that leads to substantial biases in regression coefficient estimation (see results for Settings 1-6 in Figure 4.2(a)-(b)). These differences are more noticeable in Settings 4-6 where the effects on the $0 \rightarrow 2$ and $1 \rightarrow 2$ transitions differ. In Setting 6 the true survival functions for different values of X_1 cross which clearly indicates a violation of the proportional hazards assumption. Since the Cox model estimates survival under the proportional hazards framework, this results in a substantial discrepancy between the true survival curves and those estimated by the Cox model across values of X_1 , as shown in Figure 4.6 (f).

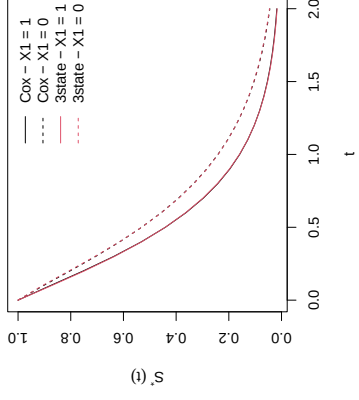
The expected Brier score for both models follows a general trend across all six settings: it begins at zero, rises over time to reach a peak near the point where the survival function is approximately 50%, and then declines. However, from Figure 4.7, it is clear that none of these six settings causes substantial impact on differences between the expected Brier score for Cox regression and that based on the illness-death model, even for Setting 6. This indicates that as the sample size increases, the difference between the estimated survival



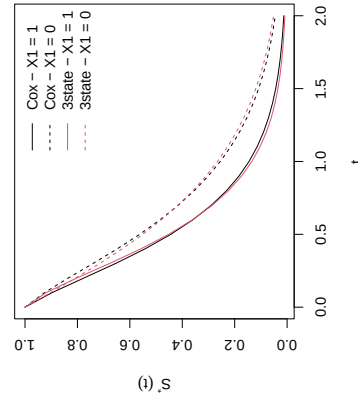
(a) Setting 1: $\beta_{01} = 0, \beta_{02} = \log 1.5, \beta_{12} = \log 1.5$



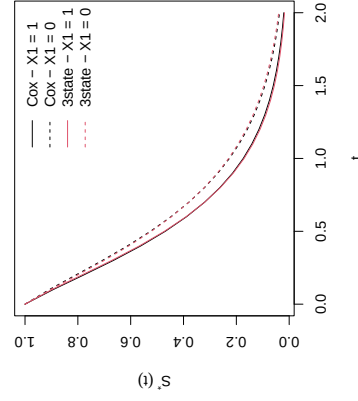
(b) Setting 2: $\beta_{01} = \log 1.5, \beta_{02} = \log 1.5, \beta_{12} = \log 1.5$



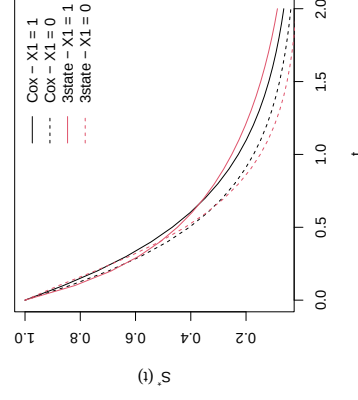
(c) Setting 3: $\beta_{01} = \log 0.5, \beta_{02} = \log 1.5, \beta_{12} = \log 1.5$



(d) Setting 4: $\beta_{01} = \log 2, \beta_{02} = \log 0.5, \beta_{12} = \log 2$

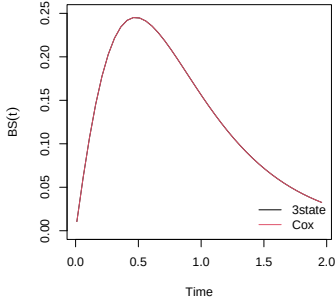


(e) Setting 5: $\beta_{01} = \log 0.5, \beta_{02} = 0, \beta_{12} = \log 2$

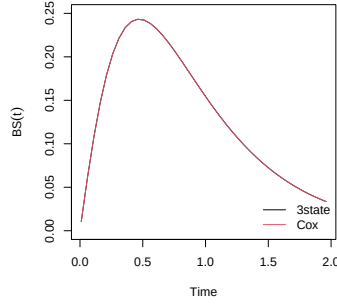


(f) Setting 6: $\beta_{01} = \log 0.5, \beta_{02} = \log 2, \beta_{12} = \log 0.5$

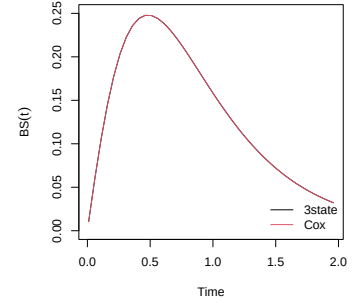
Figure 4.6: Limiting values of estimated survival functions by X_1 for Cox regression (in black) and three-state model (in red) in Settings 1–6 when $r = \lambda_{12}/\lambda_{02} = 2$.



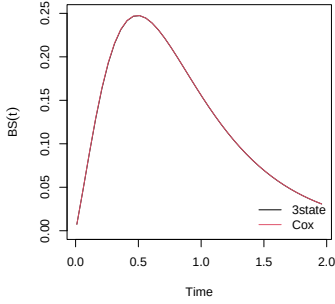
(a) Setting 1: $\beta_{01} = 0$, $\beta_{02} = \log 1.5$, $\beta_{12} = \log 1.5$



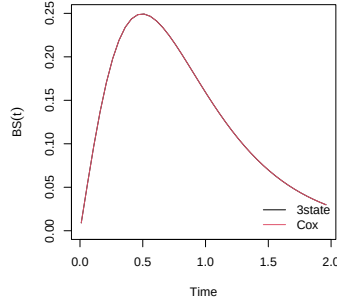
(b) Setting 2: $\beta_{01} = \log 1.5$, $\beta_{02} = \log 1.5$, $\beta_{12} = \log 1.5$



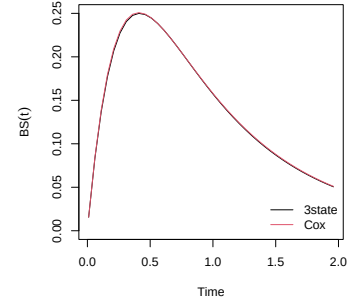
(c) Setting 3: $\beta_{01} = \log 0.5$, $\beta_{02} = \log 1.5$, $\beta_{12} = \log 1.5$



(d) Setting 4: $\beta_{01} = \log 2$, $\beta_{02} = \log 0.5$, $\beta_{12} = \log 2$



(e) Setting 5: $\beta_{01} = \log 0.5$, $\beta_{02} = 0$, $\beta_{12} = \log 2$



(f) Setting 6: $\beta_{01} = \log 0.5$, $\beta_{02} = \log 2$, $\beta_{12} = \log 0.5$

Figure 4.7: Expected Brier Score evaluated at the limiting value of estimated survival functions derived from a fitted Cox regression (in black) and three-state model (in red) in Settings 1–6 when $r = \lambda_{12}/\lambda_{02} = 2$.

function obtained from Cox regression and the true survival function accounts for only a small portion of the Brier Score, with the main component of the Brier Score arising from the irreducible error.

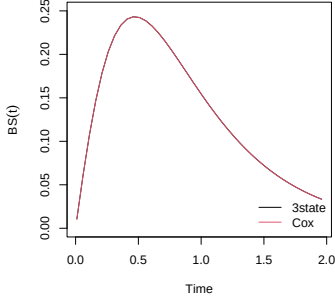
Prediction with Multiple Binary Covariates

We now consider scenarios with multiple binary covariates having heterogeneous effects on transition intensities. The settings we consider are summarized in Table 4.2. The covariates are assumed to be independent random variables from Bernoulli(0.5). Since it would be cumbersome to present the predicted conditional survival functions given all possible covariate combinations, we focus on presenting the expected Brier score when $r = 2$.

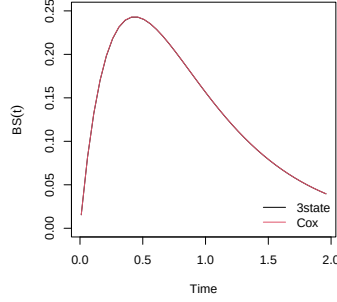
		x_1	x_2	x_3	x_4	x_5
Setting M1	β_{01}	log 1.5	log 1.5	log 1.5	log 1.5	log 1.5
	β_{02}	log 1.5	log 1.5	log 1.5	log 1.5	log 1.5
	β_{12}	log 1.5	log 1.5	log 1.5	log 1.5	log 1.5
Setting M2	β_{01}	log 1.5	log 1.5	log 1.2	log 1.2	log 1.2
	β_{02}	log 1.2	log 1.2	log 1.5	log 1.5	log 1.5
	β_{12}	log 1.1	log 1.1	log 1.1	log 1.1	log 1.1
Setting M3	β_{01}	log 2	log 0.5	log 0.5	log 0.5	log 0.5
	β_{02}	log 0.5	0	0	log 2	log 2
	β_{12}	log 2	log 2	log 2	log 0.5	log 0.5

Table 4.2: Covariate coefficients (β) for different transitions across settings and five independent covariates x_1 – x_5 generated from Bernoulli (0.5). $r = \lambda_{12}/\lambda_{02} = 2$.

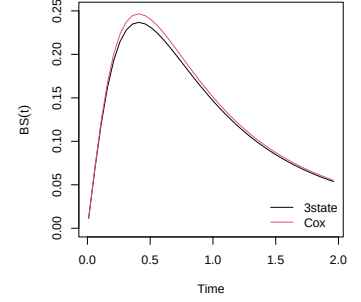
We observed slightly larger differences in the expected Brier score in Figure 4.8 compared to the scenario with only a single binary covariate.



(a) Setting M1



(b) Setting M2



(c) Setting M3

Figure 4.8: Expected Brier Score evaluated at the limiting value of estimated survival functions derived from a fitted Cox regression (in black) and three-state model (in red) in Settings M1, M2, and M3 when $r = \lambda_{12}/\lambda_{02} = 2$.

4.4.2 Omission of an Important Covariate

We consider the three settings in Figure 4.3 with the homogeneous effect of X_2 set to an extreme value $\log 0.25$ and omitted, which are referred to as Settings O1, O2, and O3, respectively. Figure 4.9 summarizes the Expected Brier Score comparing the Markov illness-death model and the Cox regression when X_2 is omitted. There is a substantial increase in the expected Brier score when the important covariate X_2 is omitted, compared to the true model. However, there is almost no difference between the expected Brier scores of the Cox model and the Markov illness-death model under omission of X_2 , with their Brier score curves nearly overlapping across different levels of association between the two covariates.

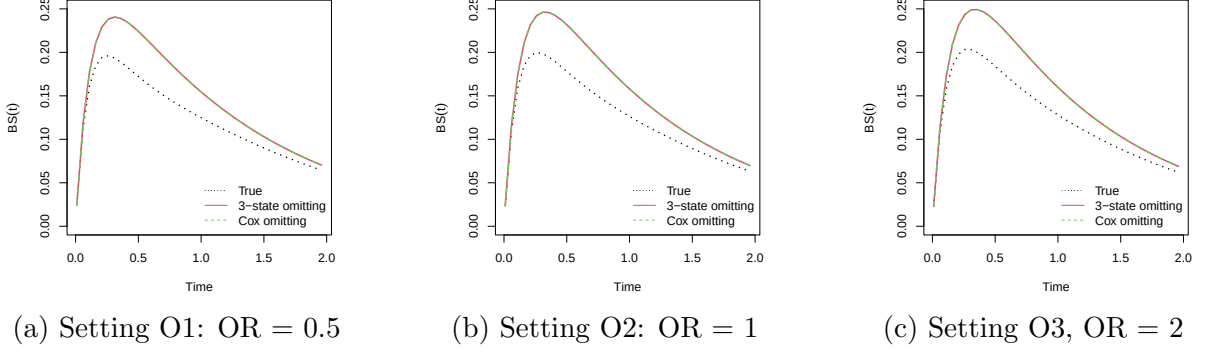


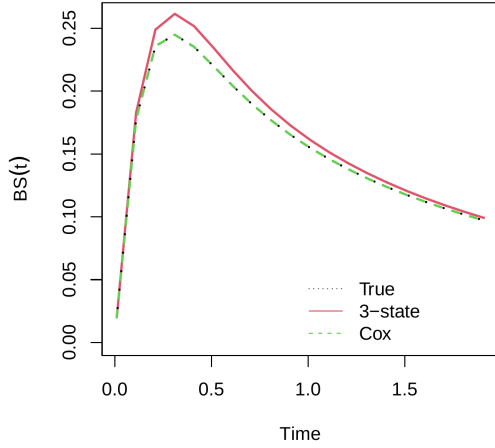
Figure 4.9: Expected Brier Score evaluated at the limiting value of survival functions derived from the true model (black dotted line), the fitted Markov illness-death model (red dashed line), and the fitted Cox regression (green solid line) when X_2 is omitted from the model. X_1 has homogeneous effect of $\log 1.5$ across all three transitions, X_2 has homogeneous effect of $\log 0.25$ across all three transitions, and $r = \lambda_{12}/\lambda_{02} = 1$. The three panels correspond to the odds ratio between X_1 and X_2 with $OR = 0.5, 1$ and 2 .

4.4.3 Violation of the Markov Assumption

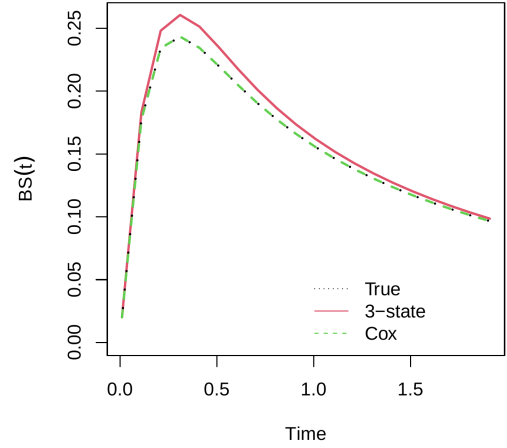
Figure 4.10 summarizes the expected Brier score comparing the Markov illness-death model and the Cox regression when the Markov assumption is violated. We examine Settings 11 and 12 in Figure 4.4 with κ set to 0.5 and $r = \lambda_{12}/\lambda_{02} = 1$. In these two settings, the effects on $0 \rightarrow 2$ and $1 \rightarrow 2$ transitions are the same, and the Cox model performs almost as well as the illness-death semi-Markov model so the expected Brier scores overlap. There is a substantial increase in the expected Brier score when the Markov illness-death model is applied, compared to the true model for both Settings 11 and 12.

4.4.4 Prediction for Disease-Progression

Here we investigate the use of the Fine-Gray model when the true data-generating process follows a Markov illness-death model. [Bühler and others \(2024\)](#) examined the limiting val-



(a) Setting 11: $\beta_{01} = 0, \beta_{02} = \beta_{12} = \log 1.5$



(b) Setting 12: $\beta_{01} = \beta_{02} = \beta_{12} = \log 1.5$

Figure 4.10: Expected Brier Score derived from the true model (in black), the fitted Markov illness-death model (in red), and the fitted Cox regression (in green) when the true process follows an illness-death semi-Markov model. The two panels correspond to Setting 1 and Setting 2, respectively, with $\kappa_{12} = 0.5$ and $r = \lambda_{12}/\lambda_{02} = 1$.

ues of regression coefficient estimators from the Fine-Gray model when data are from a randomized trial and are generated according to a competing risks process with proportional cause-specific hazards. The following derivations are similar in specification with a focus on prediction. We examine whether the Fine-Gray model can yield accurate estimates of the cumulative incidence function (CIF) and assess predictive accuracy using a Brier score adapted for competing risks ([Schoop and others, 2011](#)). Define $T_1 = \inf\{t > 0 : Z(t) = 1\}$ as the state 1 entry time, which takes the value of infinity if a $0 \rightarrow 2$ transition is made.

The cumulative incidence function of T_1 is defined as

$$F(t | X) = P(T_1 \leq t | X) = \int_0^t \exp[-\Lambda_{01}(0, u | X) - \Lambda_{02}(0, u | X)] \lambda_{01}(u | X) du. \quad (4.23)$$

We consider the Fine-Gray model of the form

$$\log(-\log(1 - F(t | X))) = \alpha(t) + \psi'_{\text{FG}} X \quad (4.24)$$

with the form of $\alpha(t)$ unspecified, which can also be characterized by a proportional hazards model for T_1 with the baseline cumulative subdistribution hazard function $\Gamma_{\text{FG}}(t) = \exp(\alpha(t))$. Let $Y_1^\dagger(s) = I(s < T_1)$ be the Fine-Gray at-risk indicator. Let $dN_1(s)$ indicate whether a $0 \rightarrow 1$ transition is made at time s .

We consider a sample of n independent but identically distributed processes and introduce a subscript i to distinguish individual processes. The Fine-Gray estimation functions are given by (Fine and Gray, 1999)

$$\sum_{i=1}^n w_i(t) Y_{i1}^\dagger(t) [dN_{i1}(t) - d\Gamma_{\text{FG}}(t) \exp(X_i' \psi_{\text{FG}})], \quad (4.25)$$

and

$$\sum_{i=1}^n \int_0^\infty w_i(t) Y_{i1}^\dagger(t) [dN_{i1}(t) - d\Gamma_{\text{FG}}(t) \exp(X_i' \psi_{\text{FG}})] X_i, \quad (4.26)$$

for $d\Gamma_{\text{FG}}(t)$ and ψ_{FG} , respectively, where the time-varying weights $w_i(t)$ are defined as

$$w_i(t) = \frac{I(C_i > \min(T_{i0}, t))}{G(\min(T_{i0}, t) | X_i)}$$

with $G(t | X_i) = P(C_i > t | X_i)$ and $T_{i0} = \inf\{t \geq 0 : Z_i(t) \neq 0\}$ is the sojourn time in

state 0. Solving (4.25) gives the “profile” estimator

$$d\widehat{\Gamma}_{\text{FG}}(t; \psi_{\text{FG}}) = \frac{\sum_{i=1}^n w_i(t) Y_{i1}^\dagger(t) dN_{i1}(t)}{\sum_{i=1}^n w_i(t) Y_{i1}^\dagger(t) \exp(X_i' \psi_{\text{FG}})}, \quad (4.27)$$

and by substituting it into (4.26) we obtain the partial score vector

$$U_{FG}(\psi_{\text{FG}}) = \sum_{i=1}^n \int_0^\infty w_i(t) Y_{i1}^\dagger(t) \left[X_i - \frac{R_{\text{FG}}^{(1)}(t; \psi_{\text{FG}})}{R_{\text{FG}}^{(0)}(t; \psi_{\text{FG}})} \right] dN_{i1}(t), \quad (4.28)$$

where $R_{\text{FG}}^{(l)}(t; \psi) = \sum_{i=1}^n w_i(t) Y_{i1}^\dagger(t) X_i^l \exp(X_i' \psi)$ for $l = 0, 1$. Given the censoring time is independent of the event process conditional on X and $G(t | X)$ assumed to be correctly specified, the limiting value of $\widehat{\psi}_{\text{FG}}$, denoted as ψ_{FG}^* , solves

$$\int_0^\infty \left[r_{\text{FG}}^{(1)}(t) - \frac{r_{\text{FG}}^{(1)}(t; \psi_{\text{FG}})}{r_{\text{FG}}^{(0)}(t; \psi_{\text{FG}})} r_{\text{FG}}^{(0)}(t) \right] dt = 0 \quad (4.29)$$

where $r_{\text{FG}}^{(l)}(t; \psi) = E \left[w(t) Y_1^\dagger(t) X^l \exp(X' \psi) \right]$ and $r_{\text{FG}}^{(l)}(t) = E \left[w(t) Y_1^\dagger(t) X^l dN_1(t) \right]$ for $l = 0, 1$. Derivations of $r_{\text{FG}}^{(l)}(t; \psi)$ and $r_{\text{FG}}^{(l)}(t)$ are given in [Bühler and others \(2024\)](#). The limiting value for $\widehat{F}_{\text{FG}}(t | X)$ from a fitted Fine-Gray model is

$$F_{\text{FG}}^*(t | X) = 1 - \exp \left[- \int_0^t d\Gamma_{\text{FG}}^*(s) \exp(X' \psi_{\text{FG}}^*) \right], \quad (4.30)$$

where

$$d\Gamma_{FG}^*(s) = \frac{r_{\text{FG}}^{(0)}(s)}{r_{\text{FG}}^{(0)}(s; \psi_{\text{FG}}^*)},$$

and ψ_{FG}^* is the limiting value for the estimated regression coefficient ψ_{FG} in the Fine-Gray model.

The expected Brier score for competing risks analysis ([Schoop and others, 2011](#)) is

defined analogously to that in survival analysis, but quantifies the discrepancy between the estimated cumulative incidence function and the observed case-specific event. Let $\hat{F}_n$ denote the limiting value of the cumulative incidence function estimated based on a sample of n independent observations, and let F^* denote the limit of $\hat{F}_n$. The expected Brier score is defined as

$$\begin{aligned} BS_{CR}(t) &= \mathbb{E} [F^*(t | X) - I(T_1 < t | X)]^2 \\ &= \mathbb{E} [F^*(t | X) - F(t | X)]^2 + E [F(t | X)(1 - F(t | X))] \end{aligned} \quad (4.31)$$

where the first term $\mathbb{E} [F^*(t | X) - F(t | X)]^2$ is referred to as imprecision, and the second term $\mathbb{E} [F(t | X)(1 - F(t | X))]$ is again called inseparability.

For a Markov illness-death model and using the 4-state presentation in Figure 4.1(b), the cumulative incidence function for $0 \rightarrow 1$ transitions can be estimated by $\hat{F}_{AJ}(t | X) = \hat{P}(Z(t) = 1 | X) + \hat{P}(Z(t) = 3 | X)$. Both $\hat{P}(Z(t) = 1 | X)$ and $\hat{P}(Z(t) = 3 | X)$ can be obtained from the estimated transition probability matrix $\hat{\mathbb{P}}(0, t | X)$ in (4.21). We consider three representative settings summarized in Table 4.3. We consider a single binary covariate, $X_1 \sim \text{Bernoulli}(0.5)$, and fix the transition coefficient $\beta_{12} = \log 1.5$ across all three settings. For the transitions $0 \rightarrow 1$ and $0 \rightarrow 2$, we examine three scenarios: (1) both effects are equal and positive, $\beta_{01} = \beta_{02} = \log 1.5$; (2) the effects are in opposite directions, $\beta_{01} = \log 2$, $\beta_{02} = \log 0.5$; and (3) both effects are negative but heterogeneous, $\beta_{01} = \log 0.5$, $\beta_{02} = \log 0.1$, corresponding to Settings C1, C2, and C3, respectively. For consistency, we maintain a 20% overall censoring rate as we did for all previous settings. We consider different probabilities of the $0 \rightarrow 1$ transitions $E_X[P(T_{01} < T_{02} | X)] = 0.3, 0.5$, and 0.7.

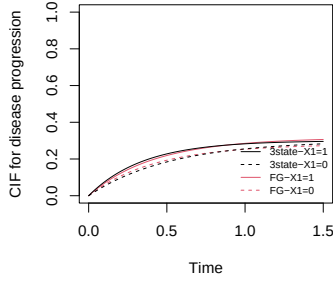
Figure 4.11 presents the limiting values of the estimated cumulative incidence functions

Coefficient	Setting C1	Setting C1	Setting C3
β_{01}	log 1.5	log 2	log 0.5
β_{02}	log 1.5	log 0.5	log 0.1
β_{12}	log 1.5	log 1.5	log 1.5

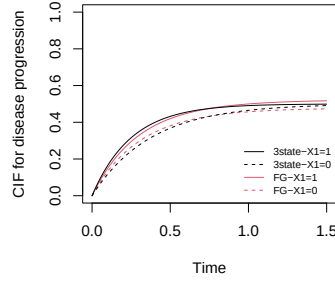
Table 4.3: Coefficients for three settings involving a single binary covariate X_1 , $\lambda_{12}/\lambda_{02} = r = 1$, and $X_1 \sim \text{Bernoulli}(0.5)$.

for the Fine-Gray model and the correctly specified Markov illness-death model by X_1 in Settings C1, C2 and C3, considering different probabilities of taking $0 \rightarrow 1$ transitions among subjects who leave state 0. In Settings C1 and C2, We observe a minor difference between the limiting values of the estimated $F(t | X)$ derived from the two models. Overall, the estimates for the cumulative incidence function from the Fine-Gray model closely track the true CIF of T_1 , $F(t | X)$. In Setting 3, the two cumulative incidence functions for $X_1 = 0$ and $X_1 = 1$ cross, whereas the Fine-Gray model assumes a proportional hazard for the subdistribution of T_1 , which leads to larger discrepancies between the limiting values of the CIFs estimated by the Fine-Gray model and the true CIFs. Figure 4.12 presents the expected Brier score changing over time in Settings C1-C3. Unlike the Brier score in survival analysis, the Brier score in competing risk analysis does not continue to decrease over time but instead levels off at a non-zero value. This behavior arises from the nature of competing risks scenarios where the proportion of subjects experiencing the event of interest never reaches one due to the presence of competing events. Another observed pattern is that, in the presence of competing events, the Brier score increases and then levels off if the probability of experiencing the event of interest is less than 0.5. If this probability is greater than 0.5, the Brier score initially increases, then decreases and plateaus. This occurs because when the probability is below 0.5, the cumulative incidence function (CIF) would never reach 0.5 which maximizes the irreducible component of the

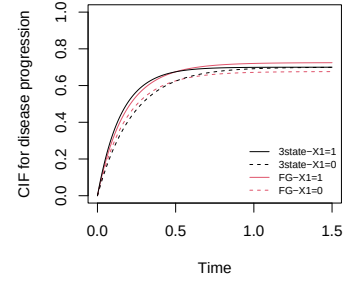
Brier score. Among Settings C1-C3, Setting C3 leads to the largest discrepancy between the Fine-Gray model and the true model, however, the difference is minor compared to the irreducible part.



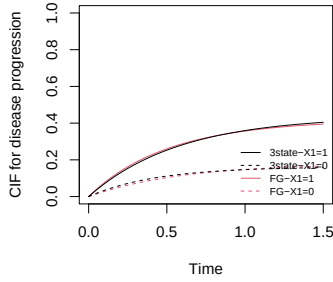
(a) Setting C1, $p_1 = 0.3$



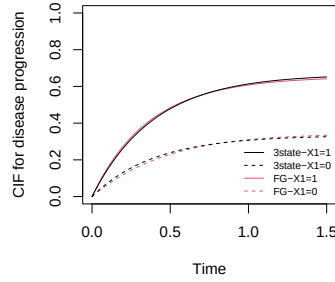
(b) Setting C1, $p_1 = 0.5$



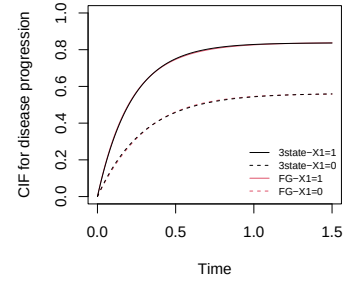
(c) Setting C1, $p_1 = 0.7$



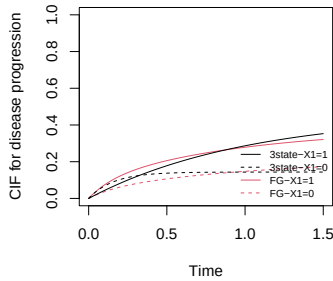
(d) Setting C2, $p_1 = 0.3$



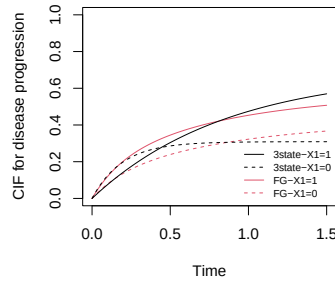
(e) Setting C2, $p_1 = 0.5$



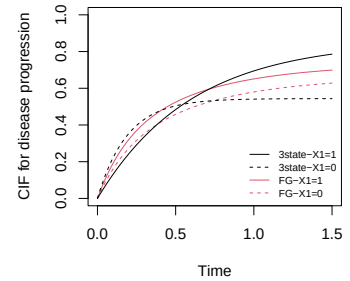
(f) Setting C2, $p_1 = 0.7$



(g) Setting C3, $p_1 = 0.3$



(h) Setting C3, $p_1 = 0.5$



(i) Setting C3, $p_1 = 0.7$

Figure 4.11: Cumulative incidence functions $F(t)$ in Setting C1, C2, and C3 with $r = \lambda_{12}/\lambda_{02} = 1$, and $p_1 = E_X[P(T_{01} < T_{02} | X)] = 0.3, 0.5$ and 0.7 . Black curves represent Fine-Gray estimates, and red curves represent true Markov illness-death model

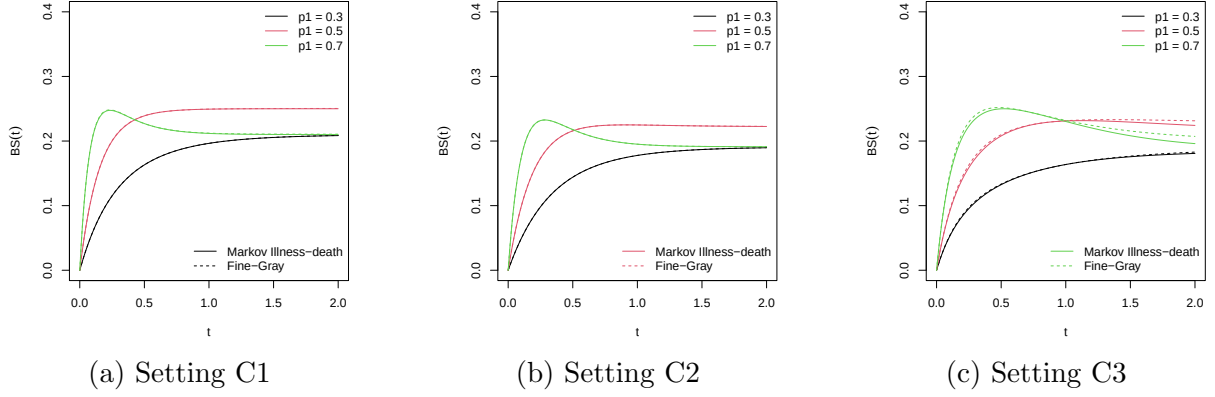


Figure 4.12: Limiting value of the expected Brier Score $BS(t)$ in Settings C1, C2, and C3 and $p_1 = E_X[P(T_{01} < T_{02} | X)] = 0.3, 0.5$, and 0.7 . Solid curves represent Fine-Gray estimates, and dashed curves represent true Markov illness-death model

4.5 Simulation Studies

We assume the covariate vector X follows a p -dimensional multivariate normal distribution $MVN_p(0, \Sigma)$, where $p = 5$ and Σ takes an autoregressive dependence structure with (j, k) entry being $\Sigma_{jk} = \kappa^{|j-k|}$, $j, k = 1, \dots, 5$, and $\kappa = 0.5$. The coefficients are set to the same as Settings M1, M2, and M3 in Table 4.2. which are previously used to investigate with five binary covariates. However, here we consider continuous covariates which are correlated and deriving the limiting values could be challenging. In Setting M1, all covariates have the same positive effects of $\log 1.5$ across all three transitions. In Setting M2, the five covariates have effects on the three transitions in the same direction, but in different sizes. In Setting M3, the effects on the $0 \rightarrow 2$ and $1 \rightarrow 2$ transitions are either in different directions or one of them is zero. We generated 100 independent processes and report the average estimates for regression coefficients for $X_1, \dots, X_5$, empirical standard error, and average robust standard error in Table 4.4.

In terms of covariate effect estimation, the Markov illness-death model produces es-

timates for covariate effects close to the true values across all three settings. The Cox regression for overall survival yields estimators that are not consistent with the true effects across three transitions. In Setting M1, the Cox regression estimator slightly overestimates the conceptualized covariate effects on the two transitions to death, which is $\log 1.5$. This is due to all covariates having positive effects of $\log(1.5)$ on the $0 \rightarrow 1$ transition, combined with an increased risk of death from the intermediate state 1 ($r = \lambda_{12}/\lambda_{02} = 1.2$). This observation aligns with the plots of Setting M2 in Figure 4.2(a), where $r > 1$ leads to an overestimation of the covariate effect on the overall risk of death. In Setting M2, the true covariate effects differ across the three transitions but remain in the same direction. The Cox regression estimates lie between the largest and smallest of these transition-specific effects for each covariate. In Setting M3, the Cox model yields covariate effects that differ substantially from the true effects on each transition. For all models, the average robust standard error tracks the empirical standard error well.

With a sample of n subjects with right-censored data, and assuming that $\widehat{G}(t)$ is a consistent estimator for the censoring distribution $G(t)$, the Brier score for predicting overall survival can be estimated by

$$\widehat{BS}(t) = \frac{1}{n} \sum_{i=1}^n \left\{ I(v_i \leq t, \delta_i = 1) \frac{1}{\widehat{G}(v_i)} \left[0 - \widehat{\mathcal{F}}(t | x_i) \right]^2 + I(v_i > t) \frac{1}{\widehat{G}(t)} \left[1 - \widehat{\mathcal{F}}(t | x_i) \right]^2 \right\}, \quad (4.32)$$

which is a consistent estimator for $BS(t)$ using the inverse probability censoring weighting (IPCW) approach (Graf and others, 1999). Similarly, the IPCW approach can be applied to competing risk analysis to construct a consistent estimator for the corresponding Brier score (Schoop and others, 2011). Define $T_0 = \inf\{t \geq 0 : Z(t) \neq 0\}$ as the sojourn time in state 0, let $V_0 = \min\{T_0, C\}$ denote the observed time, and let $\Delta_0 = I(T_0 < C)$ indicate whether it is an observed transition time out of state 0. The consistent inverse probability

censoring weighted estimator for the Brier score in competing risk analysis is given by

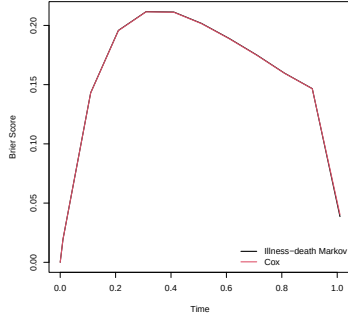
$$\begin{aligned}\widehat{BS}_{CR}(t) = & \frac{1}{n} \sum_{i=1}^n \left\{ I(v_{i0} \leq t, \delta_{i0} = 1, Z(t_0) = 1) \frac{1}{\widehat{G}(v_{i0})} \left[1 - \widehat{F}_{FG}(t \mid x_i) \right]^2 \right. \\ & + I(v_{i0} \leq t, \delta_{i0} = 1, Z(t_{i0}) = 2) \frac{1}{\widehat{G}(v_{i0})} \left[0 - \widehat{F}_{FG}(t \mid x_i) \right]^2 \\ & \left. + I(v_{i0} > t) \frac{1}{\widehat{G}(t)} \left[0 - \widehat{F}_{FG}(t \mid x_i) \right]^2 \right\}.\end{aligned}\tag{4.33}$$

For both $\widehat{BS}(t)$ and $\widehat{BS}_{CR}(t)$, we adopt the Kaplan-Meier estimator for the censoring distribution $G(t) = P(C_i \geq t)$.

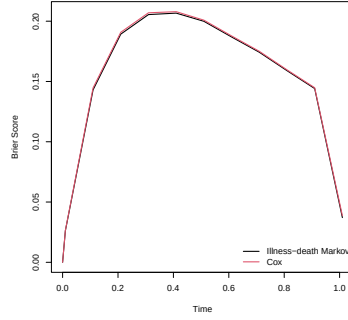
Based on the estimated survival function for the Cox regression model and the estimated cumulative incidence function for the Fine-Gray model under different settings, we calculate the empirical Brier score on an independent test set of 200 subjects at a grid of time over $(0, 1]$ with points every 0.1 units. We present the results in Figure 4.13 and Figure 4.14. We observe minimal differences between the Brier scores of the fitted Cox regression and the 3-state model for predicting overall survival in Settings M1 and M2 in Figure 4.13. However, the discrepancy becomes much more pronounced in Setting M3, where the covariate effects are set in opposite directions on the $0 \rightarrow 2$ and $1 \rightarrow 2$ transitions. Across all three settings, the differences in the Brier scores between the Fine-Gray model and the Markov illness-death model remain relatively small in Figure 4.14.

Setting	Transition	Model		X_1	X_2	X_3	X_4	X_5
M1	$0 \rightarrow 1$	True value		0.693	0.693	0.693	0.693	0.693
		3-State	EST	0.185	0.186	0.182	0.185	0.182
			ESE	0.050	0.048	0.048	0.050	0.042
			RSE	0.045	0.050	0.050	0.050	0.045
	$0 \rightarrow 2$	True value		0.693	0.693	0.693	0.693	0.693
		3-State	EST	0.175	0.187	0.187	0.180	0.183
			ESE	0.068	0.076	0.080	0.076	0.074
			RSE	0.069	0.076	0.076	0.077	0.068
	$1 \rightarrow 2$	True value		0.693	0.693	0.693	0.693	0.693
		3-State	EST	0.189	0.174	0.192	0.179	0.177
			ESE	0.053	0.055	0.053	0.056	0.053
			RSE	0.053	0.058	0.058	0.058	0.053
		Cox - OS	EST	0.191	0.189	0.197	0.187	0.187
			ESE	0.045	0.048	0.048	0.046	0.046
			RSE	0.041	0.046	0.046	0.046	0.041
		Fine-Gray	EST	0.068	0.062	0.060	0.065	0.062
			ESE	0.047	0.054	0.046	0.049	0.047
			RSE	0.046	0.051	0.051	0.052	0.046
M2	$0 \rightarrow 1$	True value		0.405	0.405	0.182	0.182	0.182
		3-State	EST	0.409	0.408	0.181	0.186	0.182
			ESE	0.049	0.047	0.046	0.050	0.043
			RSE	0.047	0.051	0.051	0.051	0.046
	$0 \rightarrow 2$	True value		0.182	0.182	0.405	0.405	0.405
		3-State	EST	0.176	0.190	0.411	0.399	0.406
			ESE	0.077	0.077	0.079	0.070	0.070
			RSE	0.071	0.078	0.078	0.078	0.070
	$1 \rightarrow 2$	True value		0.095	0.095	0.095	0.095	0.095
		3-State	EST	0.101	0.090	0.100	0.089	0.085
			ESE	0.050	0.059	0.058	0.051	0.052
			RSE	0.052	0.057	0.057	0.058	0.052
		Cox - OS	EST	0.137	0.134	0.215	0.202	0.204
			ESE	0.046	0.051	0.051	0.048	0.046
			RSE	0.043	0.048	0.048	0.048	0.043
		Fine-Gray	EST	0.236	0.229	-0.034	-0.024	-0.029
			ESE	0.047	0.054	0.047	0.048	0.045
			RSE	0.048	0.053	0.052	0.053	0.047
M3	$0 \rightarrow 1$	True value		0.693	-0.693	-0.693	-0.693	-0.693
		3-State	EST	0.700	-0.692	-0.704	-0.695	-0.699
			ESE	0.048	0.051	0.050	0.050	0.050
			RSE	0.048	0.052	0.053	0.054	0.048
	$0 \rightarrow 2$	True value		-0.693	0.000	0.000	0.693	0.693
		3-State	EST	-0.698	0.005	-0.005	0.701	0.692
			ESE	0.068	0.078	0.077	0.081	0.079
			RSE	0.074	0.077	0.077	0.081	0.074
	$1 \rightarrow 2$	True value		0.693	0.693	0.693	-0.693	-0.693
		3-State	EST	0.703	0.692	0.700	-0.705	-0.703
			ESE	0.056	0.063	0.058	0.061	0.062
			RSE	0.060	0.059	0.060	0.064	0.059
		Cox - OS	EST	0.089	0.246	0.250	-0.102	-0.093
			ESE	0.050	0.046	0.045	0.053	0.050
			RSE	0.047	0.045	0.045	0.051	0.048
		Fine-Gray	EST	0.735	-0.568	-0.582	-0.726	-0.730
			ESE	0.057	0.066	0.057	0.059	0.057
			RSE	0.053	0.058	0.057	0.059	0.052

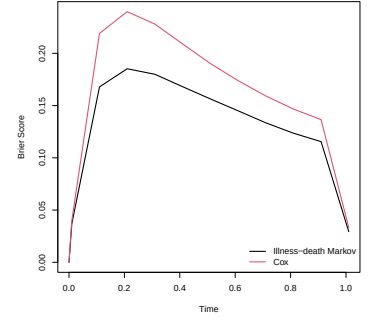
Table 4.4: Estimated coefficients (EST), empirical standard errors (ESE), and robust standard errors (RSE) when data are generated under the 3-state model with $r = \lambda_{12}/\lambda_{02} = 1.2$, and analyzed using the 3-state Markov model, Cox model, and Fine-Gray model.



(a) Setting M1

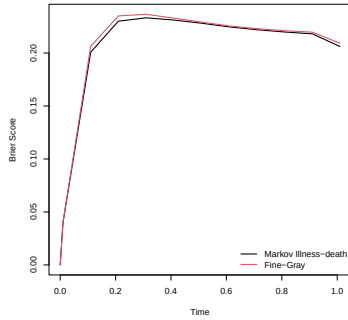


(b) Setting M2

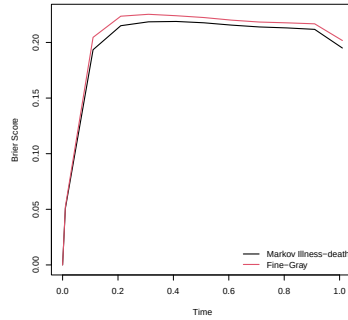


(c) Setting M3

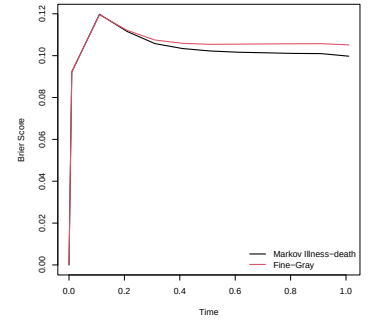
Figure 4.13: Empirical Brier score $\widehat{BS}(t)$ based on the estimated survival functions derived from the Markov illness-death model and Cox regression for overall survival, evaluated on an independent test set of 200 subjects; $r = \lambda_{12}/\lambda_{02} = 1.2$.



(a) Setting M1



(b) Setting M2



(c) Setting M3

Figure 4.14: Empirical Brier score $\widehat{BS}(t)$ based on the estimated cumulative incidence functions derived from the Markov illness-death model and Fine-Gray model for the intermediate (non-fatal) event evaluated based on an independent test set of 200 subjects; $r = \lambda_{12}/\lambda_{02} = 1.2$.

4.6 Prediction of Fractures for Patients with Metastatic Breast Cancer

Here we analyze the data from an international randomized trial evaluating the effect of pamidronate on preventing fractures in women with stage IV breast cancer metastatic to bone. Patients were followed until death or censoring. Our goal is to illustrate the relative performance of the various methods for predicting overall survival time and progression to fracture by comparing the empirical estimates of predictive accuracy of routinely applied models in practice: these include a Cox model for overall survival, an illness-death model treating fracture as an intermediate event, and a Fine-Gray model treating death as a competing risk for fracture. For simplicity, we restrict our attention to the control arm comprised of 167 patients; we use covariates including time from breast cancer diagnosis to study entry in years (`cancer.to.entry`), presence of bone metastases (`bone.mets`), bone pain score (`b.pain.score`), urinary hydroxyproline creatinine ratio in percentage (`b.hycrr`), and serum bone alkaline phosphatase (`b.alkph`). A total of 79 individuals were observed to experience a fracture, 83 subjects died, of which 14 died fracture-free. Although the censoring times for fracture and death differ, for simplicity, we use the minimum of the censoring times for fracture and death and disregard deaths occurring after censoring of the fracture.

To evaluate and compare the predictive accuracy of the three models, we randomly split the data into a training set and a test set using a 4:1 ratio. Each model was fitted on the training data under the assumption that the five covariates have multiplicative effects on the hazard or transition rate functions. Predictive performance was assessed on the test data using the Brier score for both death status by (4.32) and the occurrence of fractures by (4.33). In the three-state Markov model, we assessed both the predicted

probability of death and the probability of experiencing a fracture; the respective Brier scores could then be compared against those from the Cox regression and Fine-Gray models, respectively. The inverse probability of censoring weights used in estimating the Brier score were constructed using the Kaplan-Meier estimator for the distribution of the random censoring time.

Covariate		3-state Markov			Cox-OS	Fine-Gray
		0 → 1	0 → 2	1 → 2		
Cancer Duration (years)	EST	0.019	-0.003	-0.026	-0.006	0.023
	RSE	0.027	0.041	0.029	0.025	0.026
	p	0.484	0.939	0.368	0.825	0.382
Bone Metastases (yes or no)	EST	-0.390	0.439	-0.223	-0.070	-0.448
	RSE	0.245	0.576	0.254	0.247	0.229
	p	0.111	0.446	0.381	0.776	0.050
Bone Pain Score	EST	0.203	0.151	-0.005	0.127	0.169
	RSE	0.050	0.103	0.061	0.050	0.042
	p	< 0.001	0.143	0.931	0.011	< 0.001
HYCRR [†]	EST	-0.005	0.089	0.014	0.037	-0.018
	RSE	0.019	0.020	0.036	0.028	0.015
	p	0.815	< 0.001	0.697	0.180	0.236
Bone Alkphosph [‡]	EST	0.013	-0.048	-0.001	0.004	0.014
	RSE	0.003	0.023	0.003	0.003	0.003
	p	< 0.001	0.041	0.831	0.184	< 0.001

[†] urinary hydroxyproline/creatinine ratio (%).

[‡] bone alkaline phosphatase (units).

Table 4.5: Summary of covariate effects (EST), robust standard errors (RSE), and p-values for the 3-state Markov model, Cox model for overall survival, and Fine-Gray model for fracture.

The pain score was found to have significant effects across all three models, indicating that higher pain scores are associated with an increased risk of fracture and overall mor-

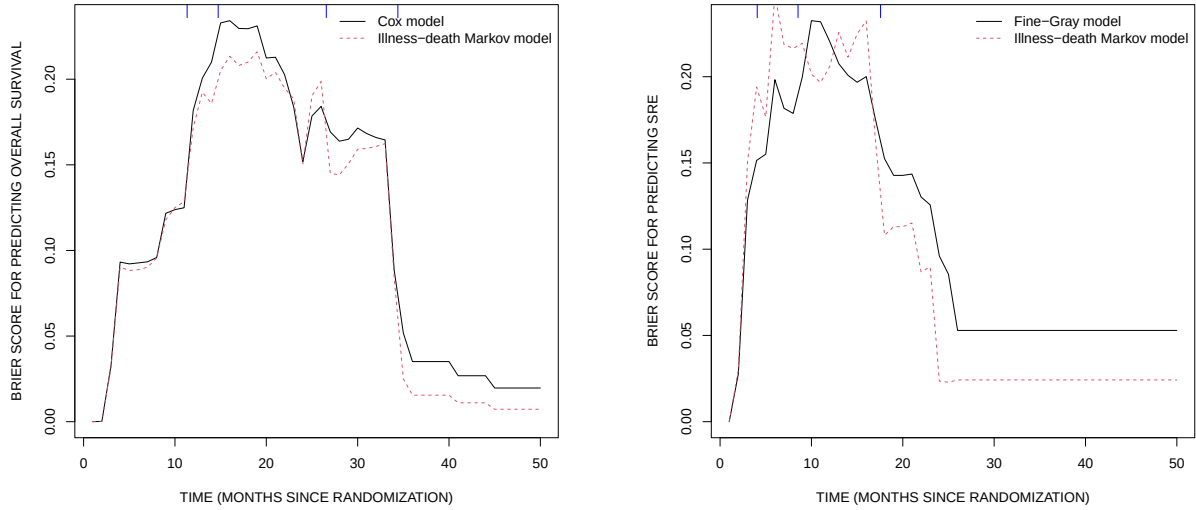


Figure 4.15: a) Brier score for predicting overall survival comparing Cox model and 3-state model, with hashmarks on the top representing the 20%, 40%, 60%, 80% quantiles of death times; b) Brier score for predicting disease progression survival comparing Fine-Gray model and 3-state model, with hashmarks on the top representing the 20%, 40%, 60% quantiles of time to fractures

tality. In the Markov illness-death model, the pain score is significant and increases the risk of experiencing a fracture ($0 \rightarrow 1$ transition), but not for the risk of death from either the fracture-free or post-fractured state. The urinary hydroxyproline creatinine ratio has a significant effect on the risk of death from the fracture-free state in a 3-state Markov model, but was not identified as significant in the Cox regression or the Fine-Gray model. Serum bone-specific alkaline phosphatase was also significant in the 3-state model which is associated with an increased risk of fracture and a lower risk of dying fracture-free; this variable was also significantly associated with fracture risk in the Fine-Gray model. Additionally, the Fine-Gray model identifies the presence of bone metastases as having a protective effect on fractures, which is not identified by the other two models. These diverse findings reflect

the fact that the models target different goals (e.g., survival or fracture), and even for a specific goal the models are generally incompatible.

Figure 4.15 presents the Brier scores for survival prediction, comparing the Markov illness-death model with Cox regression, as well as the Brier scores for fracture prediction where we compare the Markov illness-death model with the Fine-Gray model. For survival, the results show that the Markov illness-death model generally achieves better predictive accuracy than the Cox model across time, especially at the peak where around 40% to 60% subjects remain alive. This could be explained by the fact that the 3-state Markov model identifies more covariates informative for predicting time to death and accommodates different roles in the 3-state process. For the prediction of fracture occurrence, the three-state Markov model generally produced higher Brier scores than the Fine-Gray model over the first 15 months since randomization. Beyond this period, the three-state model outperformed the Fine-Gray model with a lower Brier score. Note that the variation in the estimated Brier scores has not been addressed here and these remarks are therefore qualitative - a more extensive study could offer more general conclusions.

4.7 Discussion

This work is motivated in part by the need to appropriately formulate models when conducting variable selection in chapters 2 and 3. Multistate models have been widely used to characterize processes involving multiple stages, however, in practice, simpler but often incompatible models are commonly fitted when the primary interest lies in certain marginal features. The primary goal of this research is to evaluate how model misspecifications impact the estimation of regression coefficients and predictive accuracy. We focus on two commonly used models: Cox regression and the Fine-Gray model, under the assumption

that the true underlying process follows a Markov illness-death model. Our findings indicate that the limiting values of regression coefficients estimators obtained from Cox models and Fine-Gray models depend on several factors, including the baseline hazard ratio of death between the intermediate and initial states, the probability of moving through the intermediate state, and covariate effects on all transitions. We also examine the risks associated with the use of a misspecified Markov illness-death model, by considering scenarios including omitting important covariates and violating the Markov assumption. Among the scenarios we investigated, both Markov illness-death model and Cox regression yield inconsistent estimators. Other forms of misspecification are possible. For example, failing to account for heterogeneity across patients through inclusion of random effects. In the topic of prediction, we consider scenarios with a single or multiple binary covariates since binary covariates facilitate the derivation of limiting values, but continuous covariates can also be considered. When covariates are binary, the expectations with respect to X have closed-form expressions. For continuous covariates, one could consider using Monte Carlo approximations with large datasets (e.g., $n = 10000$) to approximate the expectations with respect to covariate X . For example, the expectations involved in calculating the limiting value for the regression coefficient in Cox regression can be approximated by

$$\tilde{r}^{(0)}(s) = \frac{1}{n} \sum_{i=1}^n I(Z_i(s) = 0) \lambda_{02} \exp(x_i' \beta_{02}) + I(Z_i(s) = 1) \lambda_{12} \exp(x_i' \beta_{12}). \quad (4.34)$$

Similarly, $r^{(1)}(s)$ can be approximated by

$$\tilde{r}^{(1)}(s) = \frac{1}{n} \sum_{i=1}^n I(Z_i(s) = 0) \lambda_{02} x_i \exp(x_i' \beta_{02}) + I(Z_i(s) = 1) \lambda_{12} x_i \exp(x_i' \beta_{12}). \quad (4.35)$$

$r^{(0)}(s; \psi)$ and $r^{(1)}(s; \psi)$ are approximately

$$\tilde{r}^{(0)}(s; \psi) = \frac{1}{n} \sum_{i=1}^n I(Z_i(s) < 2) \exp(x'_i \psi), \quad (4.36)$$

and

$$\tilde{r}^{(1)}(s; \psi) = \frac{1}{n} \sum_{i=1}^n I(Z_i(s) < 2) x_i \exp(x'_i \psi), \quad (4.37)$$

respectively. The same method can be used to approximate the two marginal probabilities $P(T_{01} < T_{02})$ and $P(T_2 < \min(C, A))$ when the presence of continuous covariates makes it challenging to derive the closed forms. To solve for λ_{01} and λ_{02} such that the two constraints are met over different settings, although `nleqslv()` in *R* was previously used when closed-form expressions for the marginal probabilities were available, we suggest using the `optim()` function which is less sensitive to the stochastic variation in the target function arising from Monte Carlo approximations. Note that due to the potential rugged shape of the integrand function, the `integrate()` function in *R* could fail. In that case, one could apply the Trapezoidal Rule to evaluate the integral.

Chapter 5

Discussion and Future Research

5.1 Extensions of Variable Selection for Progressive Processes

In Chapter 2, much of the investigation was confined to simple progressive processes (with the numerical work involving a 3-state process), but the algorithm extends naturally to deal with any progressive branching process. Figure 5.1 (a) is an example of a more complex disease process which one might consider when studying the relation between the onset of degenerative disc disease and psoriatic arthritis. State 1 represents good health, while state 2 and 5 represent the presence of degenerative (spinal) disc disease (DDD), distinguishing DDD before and after PsA diagram, respectively. State 3 and 4 represent psoriatic arthritis (PsA), distinguishing PsA after and before the diagnosis of DDD, respectively. State 6 represents death. Such models are useful for studying the development of comorbidities such as these. Patients in a psoriasis clinic need to attend a clinic for examination and possible imaging, so individuals who are DDD-free and PsA-free

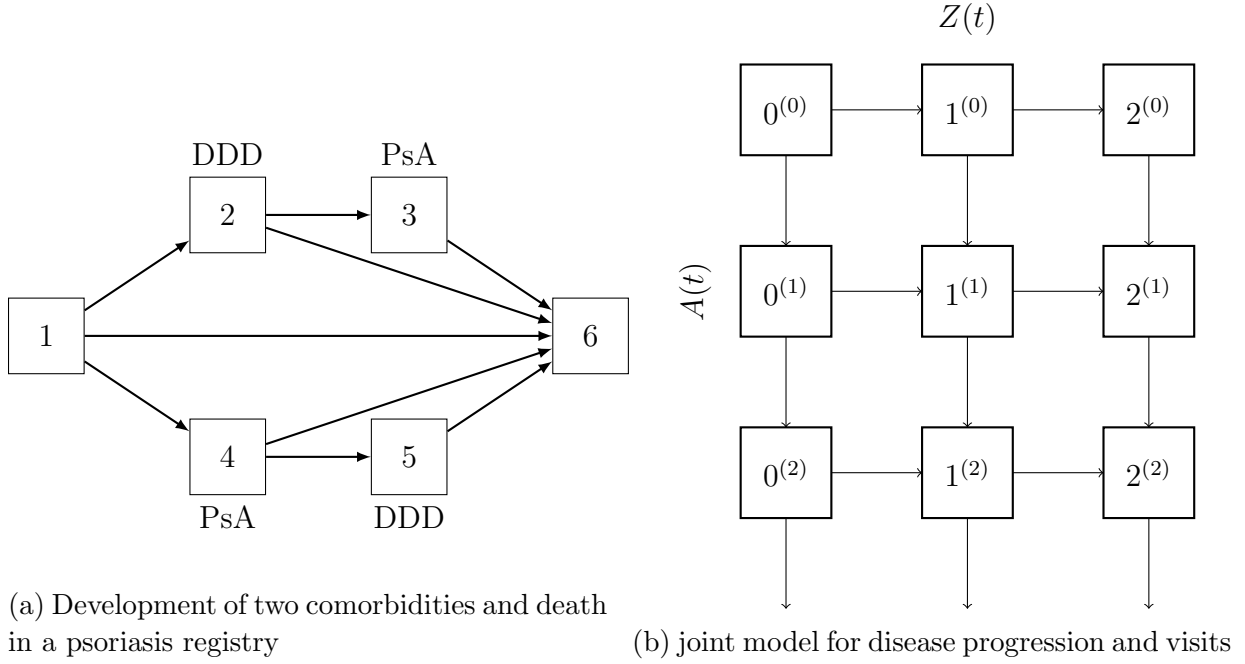


Figure 5.1: (a) A more general progressive disease process for modeling the onset of comorbidities and death and (b) A state space diagram for a joint model for a 3-state progressive process $\{Z(t), t \geq 0\}$ and a visit process $\{A(t), t \geq 0\}$

on one occasion may be found to have both PsA and DDD on the next. In this case, in addition to not knowing the time of any transitions between visits, one may not know the precise path taken between visits. In this example an individual may have made transitions $1 \rightarrow 2 \rightarrow 3$ or $1 \rightarrow 4 \rightarrow 5$. Moreover, individuals who are DDD-free and PsA-free at an assessment and then are recorded to die at the next, may have passed through one or more disease states by following paths $1 \rightarrow 2 \rightarrow 3 \rightarrow 6$, $1 \rightarrow 2 \rightarrow 6$, $1 \rightarrow 4 \rightarrow 6$, or $1 \rightarrow 4 \rightarrow 5 \rightarrow 6$, rather than the direct path $1 \rightarrow 6$ which might be naively inferred from the observed data. In such a case each possible path can be viewed as arising from a progressive process like the one depicted in Figure 2.2 and for each such process the algorithm can be applied provided an additional set of indicators is introduced labeling the

nature of the latent path. A more general complete data likelihood can then be defined and an expectation-maximization algorithm ([Dempster and others, 1977](#)) will incorporate the an expectation with respect to such indicators; given these latent indicators however, the conditional expectations for the times at risk and transition indicators for each sub-intervals in the piecewise-constant intensities can be handled as described in Chapter 2. When processes have recurrent states (i.e., processes are reversible) and are under intermittent observation, analyses are much more challenging since not knowing the paths means that the number of transitions (i.e., number of cycles through recurrent states) is unknown and there could be a countably infinite number of such paths. In this setting optimizing the penalized observed data likelihood may be more practical.

5.2 Addressing Use of Treatment

In the analysis reported on in Chapter 2, some challenges were not addressed. First, in the UTPAC cohort, many individuals received treatments which vary in strength. Nonsteroidal anti-inflammatory drugs (NSAID) are among the mildest, followed by diseasemodifying anti-rheumatic drugs (DMARDS), with biological therapies being the most powerful. Note that here we do not focus on estimating the treatment effect but rather aim to account for the treatment effect to better enable us to identify predictive markers. We consider two options for taking into account the use of treatment. One option is to include treatment as a covariate in our regression. We could consider introducing a timedependent covariate $B(t)$ to model the effect of treatment, where $B(t) = 0$ if biological therapy has no been prescribed by time t , and $B(t) = 1$ if biological therapy is prescribed by time t . Here the indicator simply indicates biologic therapy since this is by far the most powerful class of treatment. The other option is to censor the follow-up data for individuals who receive

highly effective or eradication therapy such as biological therapy at the time the treatment is introduced. These types of treatment will dramatically reduce the risk of experiencing future joint damage so one could discard follow-up observations after the date that biologic treatment starts. Because the mechanism of censoring due to receiving biological therapy is informative, the remaining data is not representative of the population of interest and one may consider using inverse probability weighting to correct for the bias.

5.3 Extensions of Variable Selection for Illness-death Models

In Chapter 3, we illustrated a penalized EM algorithm with a single LASSO penalty, two separate LASSO penalties, or a combination of LASSO and GLASSO penalties when we assume a piecewise-constant baseline for all transition intensities. However, this framework can accommodate any other available penalty functions such as the adaptive LASSO (Zou, 2006) which is of the form

$$p_{\gamma}(\boldsymbol{\beta}) = \gamma \sum_{j=0}^1 \sum_{k=j+1}^2 \sum_{l=1}^p \pi_{jkl} |\beta_{jkl}|,$$

where $\pi_{jkl} = |\tilde{\beta}_{jkl}|^{-\nu}$, $\nu \geq 1$, and $\tilde{\beta}_{jkl}$ is an estimate of β_{jkl} . In practice, $\tilde{\beta}_{jkl}$ is often taken to be the maximum likelihood estimate (MLE) and ν is set to 1 (Zhang and Lu, 2007). The smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) is another commonly used penalty, defined by

$$p_{\gamma,a}(\boldsymbol{\beta}) = \gamma \sum_{j=0}^1 \sum_{k=j+1}^2 \sum_{l=1}^p \left\{ I(|\beta_{jkl}| \leq \gamma) + \frac{(a\gamma - |\beta_{jkl}|)_+}{(a-1)\gamma} I(|\beta_{jkl}| > \gamma) \right\},$$

where $a > 2$ and $\gamma > 0$.

Note that we have shown in Section 3.3 that the function $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)})$ can be written as the summation of two components - one involves only the parameters for the 0-1 process and the other for the death process. When using a single LASSO penalty, it can be divided into two parts, with one part of the penalty applying to the conditional expectation for the 0-1 process, and the other part applying to the conditional expectation for the death process. This works because the LASSO penalty is additive with respect to the regression coefficients associated with the 0-1 process and the death process. In other words, the total LASSO penalty is the sum of the LASSO penalties for the coefficients of the 0-1 process and those of the death process. In fact, if the penalty term can be written as

$$p_\gamma(\boldsymbol{\theta}) = p_\gamma(\boldsymbol{\theta}_{01}) + p_\gamma(\boldsymbol{\theta}_2) \quad (5.1)$$

then $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)})$ can be rewritten as the sum of two penalized log-likelihood functions. Consequently, maximizing $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)})$ is equivalent to solving two separate penalized regression problems, one associated with the 0-1 process and the other involves the death process. In the previous examples of penalty functions, the LASSO, adaptive LASSO, SCAD satisfy this condition, whereas a single group LASSO does not. The advantage of being able to consider the 0-1 process and the death process separately is that it allows for greater flexibility in making assumptions about each process independently. One generalization is to maintain a piecewise assumption for the 0-1 process while adopting a flexible, unspecified form for the baseline rate of the transitions to death. This approach leads to a parametric penalized Poisson regression for the 0-1 process and a semiparametric penalized Cox regression for the death process in the M-step.

The approach we describe can be extended to accommodate more complex progressive

processes with, say, K non-terminal states where individuals in each such state can transition directly to an absorbing state; see Figure 5.2. This generalization accommodates disease processes involving multiple stages. In psoriatic arthritis, for example, patients may develop increasingly severe degrees of joint damage and progress through stages of mild disease (0 damaged joints), oligoarthritis (1 – 4 damaged joints), polyarthritis (5 – 9 damaged joints) to severe polyarthritis (≥ 10 damaged joints). In this case, states 0 to 3 can be defined based on these disease stages with state $K = 4$ representing death.

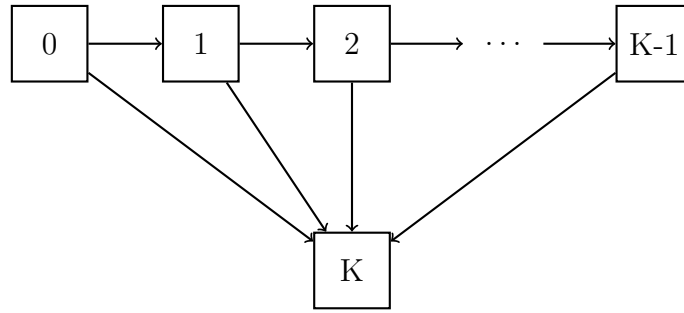


Figure 5.2: A diagram for a unidirectional progressive process with $K + 1$ states

5.4 Model Selection Criteria for Interval-censored Data

We develop methods incorporating weights into the penalty functions including the adaptive LASSO and the adaptive group LASSO where the weights change at each iteration of the EM algorithm. [Zou \(2006\)](#) proposed weights based on the ordinary least squares (OLS) estimator of effects in the context of linear regression, with maximum likelihood estimation (MLE) proposed for generalized linear models (GLMs). They also noted the oracle property holds for any $\sqrt{n}$ -consistent estimator and in cases of severe collinearity they suggest using ridge regression estimators for their stability. As noted by [Bühlmann and Van De Geer \(2011\)](#), LASSO regression estimators can also serve as initial estimators

for the ALASSO. Adaptations of the BIC for the selection of tuning parameters have also been developed for the adaptive version of the LASSO and group LASSO. The Extended Regularized Information Criterion (ERIC) has been shown to have better performance in recovering the true model over the BIC (Hui *and others*, 2015). We adopted the SABIC because this model selection criterion shows better variable selection performance than the BIC (Schwarz, 1978) when sample size is small (Chaurasia, 2023). but other procedures for selecting tuning parameters can be used.

Finally we remark that in this setting it is unclear how precisely to interpret the number of individuals in the sample (i.e., the number of independent processes) in relation to the SABIC. In one sense, the sample size under-represents the amount of information in the data since we have successive visits which, under a Markov assumption, provide more information as the number of visits increases. Moreover for a progressive 3-state process, there are two transitions per process and we acquire information regarding covariate effects from each type of transition. In another sense, the sample size could over-represent the information in our sample because of right-censoring - that is, we do not observe all individuals completing their respective processes (i.e., entering the absorbing state). Interval-censoring further reduces the information, as it means we do not actually observe any transitions. So it is challenging to think of ideal criteria for model selection in such a complex setting and this represents an important area of future research.

Similarly, in the application involving the NACC data in Chapter 3, we used the SABIC as the model selection criterion instead of BIC. When BIC was applied to the application, the model excluded all variables and thus failed to provide any useful insights. As a modified version of BIC, SABIC accommodates small sample sizes and the model selected by SABIC converges to the one selected by BIC as the sample size increases. We observed that when data is heavily censored, BIC tends to favor overly simplistic models, resulting

in a low true positive rate. The literature on appropriate model selection criteria for data subject to interval censoring and dual censoring remains sparse. Based on our experience, modifications of BIC adjusted for small samples yield better results when the effective sample size is limited. The development of variants of information criteria for incomplete time-to-event data is beyond the scope of this thesis and warrants further investigation.

5.5 High-dimensional Problems with $p > n$

Note that within the scope of our discussion, even though the number of parameters is rather large, it is still less than the sample size. With gene expression data ([Golub and others, 1999](#); [Singh and others, 2002](#)) when $p \gg n$, a practical solution is the Sure Independence Screening (SIS) ([Fan and Lv, 2008](#)). SIS reduces the dimensionality of the parameters from a very large scale to a moderate scale by ranking the features based on marginal correlations such that the number of variables is below the sample size. After the screening stage one can apply any preferred variable selection technique suitable for cases where $p < n$. This method is computationally efficient and is shown to enjoy the sure screening property, meaning that the probability that the set of variables after screening includes all important ones approaches to 1 as n goes to infinity ([Fan and Lv, 2008](#)).

5.6 Further Considerations in Model Specification and Predictive Assessment

To focus on the implications of model misspecification, we restrict our attention to the scenario where the censoring process is independent of the event process. In practice,

however, the censoring process may depend on the event process, which further affects the limiting value of the Fine-Gray estimators and the expected Brier score. The expressions for calculating the limiting values can be readily adapted to investigate the effects of dependent censoring or censoring distribution misspecification on the Fine-Gray estimator, and to evaluate the sensitivity of the IPCW-based Brier score under misspecification of the censoring distribution. Though we did not exhaustively explore all possible scenarios, the code is provided in <https://github.com/XianweiLi33/MS-misspecification> such that one could easily try other settings and calculate the limiting values.

We considered using the Brier score as the overall measure incorporating both discrimination and calibration. The integrated Brier score (Graf *and others*, 1999) is also widely used to measure the overall predictive accuracy of the predicted survival functions. Additionally, one could also investigate the implications of discrimination and calibration separately. When predicting the state at a given time horizon, the outcomes take discrete values (the states), and interest often lies in discrimination ability. Under survival models, the ability to discriminate the two states at a time horizon is typically characterized by the sensitivity, specificity, or area under the ROC curve (AUC) (Hanley and McNeil, 1982). The AUC can be interpreted as the probability that a random set of subjects from each category is correctly classified and has been extended to accommodate time-varying outcomes for assessing discrimination in survival analysis. Unlike standard binary outcomes, the outcome in survival analysis is an indicator of survival status that changes over time and the definitions of “cases” and “controls” are adapted according to study interests. Cases can be defined either as *incident* (experiencing the event exactly at time t) or *cumulative* (having experienced the event by time t), and control could be defined as *static* (subjects with the event time greater than t^* where t^* is fixed), or *dynamic* if it refers to subjects with the event time $T > t$. These definitions yield four combinations

for time-dependent ROC analysis: incident/static, incident/dynamic, cumulative/static, and cumulative/dynamic. Time-dependent sensitivity, specificity, and the area under the ROC curve (AUC) can then be derived accordingly. See [Heagerty and Zheng \(2005\)](#) for a detailed description. To address right-censoring, see the nearest neighbor estimator proposed by [Heagerty and others \(2000\)](#) and IPCW as described in [Hung and Chiang \(2010\)](#). In situations involving outcomes with more than two categories, which is the case when we adopt multistate models, the discrimination measure that can deal with polytomous outcomes is required. When there are 3 categories, the AUC in the binary situation was generalized to the Volume under the Surface (VUS) ([Mossman, 1999](#)), which can be further extended to the hypervolume under the ROC manifold (HUM) when there are more than 3 categories ([Scurfield, 1996](#)). [Van Calster and others \(2012\)](#) introduced The polytomous discrimination index (PDI), which can be computed on a category-specific basis or averaged over categories to provide a global measure. Both the VUS and PDI are defined for a set of subjects each from one of the categories, and can be interpreted similarly as the probability that a randomly selected set of subjects are correctly classified. Similar to the time-dependent AUC, calibration can be evaluated for a given sequence of time points in the context of survival analysis. A calibration plot with the estimated survival probabilities on the x-axis and the observed probabilities on the y-axis is usually used to visually assess the calibration performance ([Van Houwelingen, 2000](#)). For quantitative assessment of calibration, commonly used measures include the calibration slope ([Van Houwelingen, 2000](#)), maximum absolute calibration error ([Harrell, 2015](#)), and integrated calibration index (ICI) ([Austin and Steyerberg, 2019](#)). Further investigation is needed to understand how model misspecification influences predictive accuracy in terms of these aspects.

Bibliography

- AALEN, ODD O AND JOHANSEN, SØREN. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, 141–150.
- AKAIKE, HIROTUGU. (2003). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6), 716–723.
- ANDERSEN, PER K, BORGAN, ORNULF, GILL, RICHARD D AND KEIDING, NIELS. (2012). *Statistical Models Based on Counting Processes*. Springer Science & Business Media.
- ANDERSEN, PER KRAGH, ESBJERG, SILLE AND SØRENSEN, THORKILD I. A. (2000). Multi-state models for bleeding episodes and mortality in liver cirrhosis. *Statistics in Medicine* **19**(4), 587–599.
- ANDERSEN, PER KRAGH, HANSEN, METTE GERSTER AND KLEIN, JOHN P. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis* **10**(4), 335–350.
- ANDERSEN, PER KRAGH AND RAVN, HENRIK. (2023). *Models for Multi-State Survival Data: Rates, Risks, and Pseudo-values*. CRC Press.

- ANDERSON-BERGMAN, CLIFFORD. (2017). An efficient implementation of the EMICM algorithm for the interval censored NPMLE. *Journal of Computational and Graphical Statistics* **26**(2), 463–467.
- ARCE RENTERÍA, MIGUEL, MOBLEY, TAYLOR M, EVANGELISTA, NICOLE D, MEDINA, LUIS D, DETERS, KACIE D, FOX-FULLER, JOSHUA T, MINTO, LEX R, AVILA-RIEGER, JUSTINA AND BETTCHER, BRIANNE M. (2023). Representativeness of samples enrolled in Alzheimer’s disease research centers. *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring* **15**(2), e12450.
- AUSTIN, PETER C AND STEYERBERG, EWOUT W. (2019). The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine* **38**(21), 4051–4065.
- BEEKLY, DUANE L, RAMOS, ERIN M, LEE, WILLIAM W, DEITRICH, WOODROW D, JACKA, MARY E, WU, JOYLEE, HUBBARD, JANENE L, KOEPESELL, THOMAS D, MORRIS, JOHN C, KUKULL, WALTER A *and others*. (2007). The national Alzheimer’s coordinating center (NACC) database: the uniform data set. *Alzheimer Disease & Associated Disorders* **21**(3), 249–258.
- BERG, LEONARD. (1984). Clinical dementia rating. *The British Journal of Psychiatry* **145**(3), 339–339.
- BERNARDO, MV PATRICIA AND HARRINGTON, DAVID P. (2001). Sample size calculations for the two-sample problem using the multiplicative intensity model. *Statistics in Medicine* **20**(4), 557–579.
- BORUVKA, AUDREY AND COOK, RICHARD J. (2016). Sieve estimation in a Markov illness-death process under dual censoring. *Biostatistics* **17**(2), 350–363.

- BÜHLER, ALEXANDRA, COOK, RICHARD J AND LAWLESS, JERALD F. (2024). Estimators and cumulative incidence function regression in clinical trials: some new results on interpretability and robustness. *Statistics in Medicine* **43**(29), 5513–5533.
- BÜHLMANN, PETER AND VAN DE GEER, SARA. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.
- CHAE, KUM JU, CHOI, HYEMI, JEONG, WON GI AND KIM, JINHEUM. (2022). The value of the illness-death model for predicting outcomes in patients with non-small cell lung cancer. *Cancer Research and Treatment* **54**(4), 996–1004.
- CHAURASIA, ASHOK. (2023). Model-modified BIC as a competitor of BIC variants for model selection in regression and order selection in time series. *Communications in Statistics-Theory and Methods* **52**(23), 8425–8453.
- CHEN, JIAHUA AND CHEN, ZEHUA. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**(3), 759–771.
- COOK, RICHARD JOHN AND LAWLESS, JERALD F. (2007). *The Statistical Analysis of Recurrent Events*. Springer.
- COOK, RICHARD J AND LAWLESS, JERALD F. (2014). Statistical issues in modeling chronic disease in cohort studies. *Statistics in Biosciences* **6**, 127–161.
- COOK, RICHARD J AND LAWLESS, JERALD F. (2018). *Multistate Models for the Analysis of Life History Data*. CRC Press.
- COOK, RICHARD J AND LAWLESS, JERALD F. (2021). Independence conditions and the analysis of life history studies with intermittent observation. *Biostatistics* **22**(3), 455–481.

- COOK, RICHARD J, LAWLESS, JERALD F AND XIE, BINGFENG. (2022). Marker-dependent observation and carry-forward of internal covariates in Cox regression. *Life-time Data Analysis* **28**(4), 560–584.
- COX, DAVID R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**(2), 187–202.
- COX, DAVID R. (2017). *The Theory of Stochastic Processes*. Routledge.
- DANG, XUAN, HUANG, SHUAI AND QIAN, XIAONING. (2021). Risk factor identification in heterogeneous disease progression with l1-regularized multi-state models. *Journal of Healthcare Informatics Research* **5**, 20–53.
- DEMPSTER, ARTHUR P, LAIRD, NAN M AND RUBIN, DONALD B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **39**(1), 1–22.
- EDER, LIHI, THAVANESWARAN, ARANE, CHANDRAN, VINOD AND GLADMAN, DAFNA D. (2013). Gender difference in disease expression, radiographic damage and disability among patients with psoriatic arthritis. *Annals of the Rheumatic Diseases* **72**(4), 578–582.
- FAN, JIANQING AND LI, RUNZE. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**(456), 1348–1360.
- FAN, JIANQING AND LV, JINCHI. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(5), 849–911.

- FINE, JASON P AND GRAY, ROBERT J. (1999). A proportional hazards model for the sub-distribution of a competing risk. *Journal of the American Statistical Association* **94**(446), 496–509.
- FOUCHER, YOHANN, GIRAL, MAGALI, SOULILLOU, JEAN-PAUL AND DAURES, JEAN-PIERRE. (2007). A semi-Markov model for multistate and interval-censored data with multiple terminal events. application in renal transplantation. *Statistics in Medicine* **26**(30), 5381–5393.
- FRIEDMAN, JEROME, HASTIE, TREVOR AND TIBSHIRANI, ROB. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1), 1.
- GENANT, HK, ENGELKE, K AND PREVRHAL, S. (2008). Advanced CT bone imaging in osteoporosis. *Rheumatology* **47**(suppl_4), iv9–iv16.
- GERDS, THOMAS A, SCHEIKE, THOMAS H AND ANDERSEN, PER K. (2012). Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in Medicine* **31**(29), 3921–3930.
- GERDS, THOMAS A AND SCHUMACHER, MARTIN. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal* **48**(6), 1029–1040.
- GLADMAN, DD, FAREWELL, VT, KOPCIUK, KA AND COOK, R.J. (1998). HLA markers and progression in psoriatic arthritis. *The Journal of Rheumatology* **25**(4), 730–733.
- GLADMAN, DAFNA D AND CHANDRAN, VINOD. (2011). Observational cohort studies: lessons learnt from the University of Toronto Psoriatic Arthritis Program. *Rheumatology* **50**(1), 25–31.

- GLENN, W BRIER *and others*. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**(1), 1–3.
- GOLUB, GENE H, HEATH, MICHAEL AND WAHBA, GRACE. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**(2), 215–223.
- GOLUB, TODD R, SLONIM, DONNA K, TAMAYO, PABLO, HUARD, CHRISTINE, GAASENBEEK, MICHELLE, MESIROV, JILL P, COLLIER, HILARY, LOH, MIGNON L, DOWNING, JAMES R, CALIGIURI, MARK A *and others*. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**(5439), 531–537.
- GRAF, ERIKA, SCHMOOR, CLAUDIA, SAUERBREI, WILLI AND SCHUMACHER, MARTIN. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* **18**(17-18), 2529–2545.
- HAND, D.J. (1997). *Construction and Assessment of Classification Rules*, Wiley Series in Probability and Statistics. Wiley.
- HANLEY, JAMES A AND MCNEIL, BARBARA J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**(1), 29–36.
- HARRELL, FRANK E JR. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, Second edition. edition., Springer Series in Statistics. Cham: Springer Nature.
- HASTIE, TREVOR, TIBSHIRANI, ROBERT, FRIEDMAN, JEROME H AND FRIEDMAN, JEROME H. (2009). *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, Volume 2. Springer.

- HEAGERTY, PATRICK J, LUMLEY, THOMAS AND PEPE, MARGARET S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**(2), 337–344.
- HEAGERTY, PATRICK J AND ZHENG, YINGYE. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61**(1), 92–105.
- HOERL, ARTHUR E AND KENNARD, ROBERT W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**(1), 55–67.
- HORTOBAGYI, GABRIEL N, THERIAULT, RICHARD L, PORTER, LESTER, BLAYNEY, DOUGLAS, LIPTON, ALLAN, SINOFF, CLIVE, WHEELER, HELEN, SIMEONE, JOSEPH F, SEAMAN, JOHN, KNIGHT, ROBERT D *and others*. (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.
- HUI, FRANCIS KC, WARTON, DAVID I AND FOSTER, SCOTT D. (2015). Tuning parameter selection for the adaptive LASSO using ERIC. *Journal of the American Statistical Association* **110**(509), 262–269.
- HUNG, HUNG AND CHIANG, CHIN-TSANG. (2010). Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics* **38**(1), 8–26.
- IBRAHIM, JOSEPH G, CHU, HAITAO AND CHEN, LIDDY M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology* **28**(16), 2796–2801.
- IKUMI, KYOKO, KOBAYASHI, SHIGETO, TAMURA, NAOTO, TADA, KURISU, INOUE, HISASHI, OSAGA, SATOSHI, NISHIDA, EMI AND MORITA, AKIMICHI. (2019). HLA-

- B46 is associated with severe sacroiliitis in japanese patients with psoriatic arthritis. *Modern Rheumatology* **29**(6), 1017–1022.
- JACKSON, CHRISTOPHER. (2011). Multi-state models for panel data: the msm package for R. *Journal of Statistical Software* **38**, 1–28.
- KALBFLEISCH, JD AND LAWLESS, JERALD FRANKLIN. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* **80**(392), 863–871.
- KALBFLEISCH, JOHN D AND PRENTICE, ROSS L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.
- KANG, MINHEE AND LAGAKOS, STEPHEN W. (2007). Statistical methods for panel data from a semi-Markov process, with application to HPV. *Biostatistics* **8**(2), 252–264.
- KAPLAN, EDWARD L AND MEIER, PAUL. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**(282), 457–481.
- KEIDING, NIELS. (1991). Age-specific incidence and prevalence: a statistical perspective. *Journal of the Royal Statistical Society Series A: Statistics in Society* **154**(3), 371–396.
- LANGE, JANE M AND MININ, VLADIMIR N. (2013). Fitting and interpreting continuous-time latent Markov models for panel data. *Statistics in Medicine* **32**(26), 4581–4595.
- LAWLESS, JF 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, JERALD F AND NAZERI RAD, N. (2015). Estimation and assessment of Markov multistate models with intermittent observations on individuals. *Lifetime Data Analysis* **21**, 160–179.

- LAWLESS, JERALD F AND YUAN, YAN. (2010). Estimation of prediction error for survival models. *Statistics in Medicine* **29**(2), 262–274.
- LEFFONDRÉ, KAREN, TOURAINE, CÉLIA, HELMER, CATHERINE AND JOLY, PIERRE. (2013). Interval-censored time-to-event and competing risk with death: is the illness-death model more accurate than the Cox model? *International Journal of Epidemiology* **42**(4), 1177–1186.
- LI, CHENXI, PAK, DAEWOO AND TODEM, DAVID. (2020*a*). Adaptive LASSO for the Cox regression with interval censored and possibly left truncated data. *Statistical Methods in Medical Research* **29**(4), 1243–1255.
- LI, SHUWEI, WU, QIWEI AND SUN, JIANGUO. (2020*b*). Penalized estimation of semiparametric transformation models with interval-censored data and application to Alzheimer’s disease. *Statistical Methods in Medical Research* **29**(8), 2151–2166.
- LI, XIANWEI, COOK, RICHARD J AND DIAO, LIQUN. (2025). Variable selection for progressive multistate processes under intermittent ooservation. *Statistics in Medicine* **44**(6), e70023.
- LIN, DANYU Y AND WEI, LEE-JEN. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American statistical Association* **84**(408), 1074–1078.
- LINDEN, SJEFF VAN DER, VALKENBURG, HANS A AND CATS, ARNOLD. (1984). Evaluation of diagnostic criteria for ankylosing spondylitis. *Arthritis and Rheumatism* **27**(4), 361–368.
- LINDSEY, JANE C AND RYAN, LOUISE M. (1998). Methods for interval-censored data. *Statistics in Medicine* **17**(2), 219–238.

- LIU, LI-ZHI, WU, FANG-XIANG AND ZHANG, WEN-JUN. (2014). A group LASSO-based method for robustly inferring gene regulatory networks from multiple time-course datasets. *BMC Systems Biology* **8**, 1–12.
- MACHADO, ROBSON JM, VAN DEN HOUT, ARDO AND MARRA, GIAMPIERO. (2021). Penalised maximum likelihood estimation in multi-state models for interval-censored data. *Computational Statistics & Data Analysis* **153**, 107057.
- MEIER, LUKAS, VAN DE GEER, SARA AND BÜHLMANN, PETER. (2008). The group LASSO for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(1), 53–71.
- MEIRA-MACHADO, LUÍS AND FARIA, SUSANA. (2014). A simulation study comparing modeling approaches in an illness-death multi-state model. *Communications in Statistics-Simulation and Computation* **43**(5), 929–946.
- MILLER, RACHEL G, ANDERSON, STEWART J, COSTACOU, TINA, SEKIKAWA, AKIRA AND ORCHARD, TREVOR J. (2018). Hemoglobin a1c level and cardiovascular disease incidence in persons with type 1 diabetes: an application of joint modeling of longitudinal and time-to-event data in the pittsburgh epidemiology of diabetes complications study. *American Journal of Epidemiology* **187**(7), 1520–1529.
- MOSSMAN, DOUGLAS. (1999). Three-way ROCs. *Medical Decision Making* **19**(1), 78–89.
- NATHAN, DAVID M AND GROUP, DCCT/EDIC RESEARCH. (2014). The diabetes control and complications trial/epidemiology of diabetes interventions and complications study at 30 years: overview. *Diabetes Care* **37**(1), 9–16.

- NELDER, JOHN ASHWORTH AND WEDDERBURN, ROBERT WM. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society* **135**(3), 370–384.
- PANG, Y, KUKULL, W, SANO, M, ALBIN, RL, SHEN, C, ZHOU, J AND DODGE, HIROKO H. (2023). Predicting progression from normal to MCI and from MCI to AD using clinical variables in the national Alzheimer’s coordinating center uniform data set version 3: Application of machine learning models and a probability calculator. *The Journal of Prevention of Alzheimer’s Disease* **10**(2), 301–313.
- PERERA, STEFAN, COOK, RICHARD, LEE, KER-AI, KATZ, PATTI AND TOUMA, ZAHL. (2023). Intraindividual change in cognitive function among adults with systemic lupus erythematosus: a Markov analysis over 7 years. *ACR Open Rheumatology* **5**(3), 124–131.
- PRENTICE, R. L., KALBFLEISCH, J. D., PETERSON, A. V., FLOURNOY, N., FAREWELL, V. T. AND BRESLOW, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34**(4), 541–554.
- R CORE TEAM, R *and others*. (2013). R: A language and environment for statistical computing.
- RAHMAN, P, GLADMAN, DD, COOK, RJ, ZHOU, Y, YOUNG, G AND SALONEN, D. (1998). Radiological assessment in psoriatic arthritis. *British Journal of Rheumatology* **37**(7), 760–765.
- ROSEN, LEE S, GORDON, DAVID, KAMINSKI, MARY, HOWELL, ANTHONY, BELCH, ANDREW, MACKEY, JOHN, APFFELSTAEDT, JUSTUS, HUSSEIN, MOHAMAD A, COLEMAN, ROBERT E, REITSMA, DIRK J *and others*. (2003). Long-term efficacy and safety of zoledronic acid compared with pamidronate disodium in the treatment of

- skeletal complications in patients with advanced multiple myeloma or breast carcinoma: a randomized, double-blind, multicenter, comparative trial. *Cancer: Interdisciplinary International Journal of the American Cancer Society* **98**(8), 1735–1744.
- SAFIEH, MIRNA, KORCZYN, AMOS D AND MICHAELSON, DANIEL M. (2019). ApoE4: an emerging therapeutic target for Alzheimer’s disease. *BMC Medicine* **17**, 1–17.
- SATTEN, GLEN A AND STERNBERG, MAYA R. (1999). Fitting semi-Markov models to interval-censored data with unknown initiation times. *Biometrics* **55**(2), 507–513.
- SCHEIKE, THOMAS H, ZHANG, MEI-JIE AND GERDS, THOMAS A. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika* **95**(1), 205–220.
- SCHEMPER, MICHAEL AND HENDERSON, ROBIN. (2000). Predictive accuracy and explained variation in Cox regression. *Biometrics* **56**(1), 249–255.
- SCHOOP, ROTRAUT, BEYERSMANN, JAN, SCHUMACHER, MARTIN AND BINDER, HARALD. (2011). Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal* **53**(1), 88–112.
- SCHWARZ, GIDEON. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464.
- SCLOVE, STANLEY L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika* **52**, 333–343.
- SCURFIELD, BRIAN K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology* **40**(3), 253–269.

- SENNHENN-REULEN, HOLGER AND KNEIB, THOMAS. (2016). Structured fusion LASSO penalized multi-state models. *Statistics in Medicine* **35**(25), 4637–4659.
- SHI, YIDAN, ZENG, LEILEI, THOMPSON, MARY E AND TYAS, SUZANNE L. (2021). Augmented likelihood for incorporating auxiliary information into left-truncated data. *Lifetime Data Analysis* **27**(3), 460–480.
- SIMON, NOAH, FRIEDMAN, JEROME, HASTIE, TREVOR AND TIBSHIRANI, ROB. (2011). Regularization paths for Cox’s proportional hazards model via coordinate descent. *Journal of Statistical Software* **39**(5), 1.
- SINGH, DINESH, FEBBO, PHILLIP G, ROSS, KENNETH, JACKSON, DONALD G, MANOLA, JUDITH, LADD, CHRISTINE, TAMAYO, PABLO, RENSHAW, ANDREW A, D’AMICO, ANTHONY V, RICHIE, JEROME P *and others*. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**(2), 203–209.
- STERNBERG, MAYA R AND SATTEN, GLEN A. (1999). Discrete-time nonparametric estimation for semi-Markov models of chain-of-events data subject to interval censoring and truncation. *Biometrics* **55**(2), 514–522.
- STEYERBERG, EWOUT W. (2009). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, 1. Aufl. edition., Statistics for Biology and Health. New York, NY: Springer-Verlag.
- STONE, MERVYN. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)* **36**(2), 111–133.
- STRUTHERS, CYNTHA A AND KALBFLEISCH, JOHN D. (1986). Misspecified proportional hazard models. *Biometrika* **73**(2), 363–369.

- SUN, JIANGUO. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*, Volume 3. Springer.
- SUN, LIUQUAN, LI, SHUWEI, WANG, LIANMING, SONG, XINYUAN AND SUI, XUEMEI. (2022). Simultaneous variable selection in regression analysis of multivariate interval-censored data. *Biometrics* **78**(4), 1402–1413.
- SUTRADHAR, RINKU AND COOK, RICHARD J. (2008). Analysis of interval-censored data from clustered multistate processes: application to joint damage in psoriatic arthritis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **57**(5), 553–566.
- SUTRADHAR, RINKU AND COOK, RICHARD J. (2009). A bivariate mover–stayer model for interval-censored recurrent event data: Application to joint damage in rheumatology. *Communications in Statistics—Theory and Methods* **38**(18), 3389–3405.
- SWEETING, MJ, FAREWELL, VT AND DE ANGELIS, D. (2010). Multi-state Markov models for disease progression in the presence of informative examination times: An application to hepatitis C. *Statistics in Medicine* **29**(11), 1161–1174.
- TIAN, TIAN AND SUN, JIANGUO. (2023). Variable selection for nonparametric additive Cox model with interval-censored data. *Biometrical Journal* **65**(1), 2100310.
- TIBSHIRANI, ROBERT. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.
- TIBSHIRANI, ROBERT. (1997). The LASSO method for variable selection in the Cox model. *Statistics in Medicine* **16**(4), 385–395.
- TIBSHIRANI, ROBERT, SAUNDERS, MICHAEL, ROSSET, SAHARON, ZHU, JI AND

- KNIGHT, KEITH. (2005). Sparsity and smoothness via the fused LASSO. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(1), 91–108.
- TITMAN, ANDREW C AND SHARPLES, LINDA D. (2010). Semi-Markov models with phase-type sojourn distributions. *Biometrics* **66**(3), 742–752.
- TOURAINÉ, CÉLIA, HELMER, CATHERINE AND JOLY, PIERRE. (2016a). Predictions in an illness-death model. *Statistical Methods in Medical Research* **25**(4), 1452–1470.
- TOURAINÉ, CÉLIA, HELMER, CATHERINE AND JOLY, PIERRE. (2016b). Predictions in an illness-death model. *Statistical Methods in Medical Research* **25**(4), 1452–1470.
- TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **38**(3), 290–295.
- UNO, HAJIME, CAI, TIANXI, PENCINA, MICHAEL J, D’AGOSTINO, RALPH B AND WEI, LEE-JEN. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* **30**(10), 1105–1117.
- VAN CALSTER, BEN, VAN BELLE, VANYA, VERGOUWE, YVONNE, TIMMERMAN, DIRK, VAN HUFFEL, SABINE AND STEYERBERG, EWOUT W. (2012). Extending the C-statistic to nominal polytomous outcomes: the polytomous discrimination index. *Statistics in Medicine* **31**(23), 2610–2626.
- VAN HOUWELINGEN, HANS C. (2000). Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine* **19**(24), 3401–3415.
- VOLINSKY, CHRIS T AND RAFTERY, ADRIAN E. (2000). Bayesian information criterion for censored survival models. *Biometrics* **56**(1), 256–262.

- WANG, HANSHENG AND LENG, CHENLEI. (2008). A note on adaptive group LASSO. *Computational Statistics & Data Analysis* **52**(12), 5277–5286.
- WANG, HANSHENG, LI, RUNZE AND TSAI, CHIH-LING. (2007a). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**(3), 553–568.
- WANG, LIFENG, CHEN, GUANG AND LI, HONGZHE. (2007b). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23**(12), 1486–1494.
- WU, LONGYANG AND COOK, RICHARD J. (2012). Misspecification of Cox regression models with composite endpoints. *Statistics in Medicine* **31**(28), 3545–3562.
- WU, YING AND COOK, RICHARD J. (2015). Penalized regression for interval-censored times of disease progression: Selection of HLA markers in psoriatic arthritis. *Biometrics* **71**(3), 782–791.
- WULFSOHN, MICHAEL S AND TSIATIS, ANASTASIOS A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 330–339.
- XU, JINFENG, KALBFLEISCH, JOHN D AND TAI, BEECHOO. (2010). Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics* **66**(3), 716–725.
- XU, YANG, ZHAO, SHISHUN, HU, TAO AND SUN, JIANGUO. (2021). Variable selection for generalized odds rate mixture cure models with interval-censored failure time data. *Computational Statistics & Data Analysis* **156**, 107115.
- YUAN, MING AND LIN, YI. (2006). Model selection and estimation in regression with

- grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1), 49–67.
- ZENG, LEILEI, COOK, RICHARD J AND LEE, KER-AI. (2018). Design of cancer trials based on progression-free survival with intermittent assessment. *Statistics in Medicine* **37**(12), 1947–1959.
- ZHANG, HAO HELEN AND LU, WENBIN. (2007). Adaptive LASSO for Cox’s proportional hazards model. *Biometrika* **94**(3), 691–703.
- ZHANG, YIYUN, LI, RUNZE AND TSAI, CHIH-LING. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* **105**(489), 312–323.
- ZHAO, HUI, WU, QIWEI, LI, GANG AND SUN, JIANGUO. (2020). Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. *Journal of the American Statistical Association* **115**(529), 204–216.
- ZHONG, YUJIE AND COOK, RICHARD J. (2021). Semiparametric recurrent event vs time-to-first-event analyses in randomized trials: Estimands and model misspecification. *Statistics in Medicine* **40**(16), 3823–3842.
- ZOU, HUI. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* **101**(476), 1418–1429.

APPENDICES

Appendix A

Appendix to Chapter 2

A.1 Simulation Results with A Smaller Sample Size

Here we investigate the performance of the multistate penalized EM algorithm and naive penalized Cox regression following ad hoc right-endpoint or mid-point imputation with a smaller sample size. Table [A.1](#) and Table [A.2](#) show the empirical transition-specific and covariate-wise true positive and false positive rates (as percentages) for the multistate analysis and the Cox regression when sample size $n = 500$ with $\rho = 16$ or 4 as the mean number of visits over $(0, 1]$. The results show a degradation in performance compared to when $n = 1000$; see Tables [2.1](#) and [2.2](#). Specifically there is a drop in the true positive and increase in the false positive rate. As noted in the main paper the maximum number of iterations of the EM algorithm was set to 100 and with the smaller sample size of $n = 500$ the algorithm did not always fully converge by the final iteration. Specifically when $n = 500$ and $\rho = 4$ the proportion of datasets where the EM algorithm fails to converge within 100 iterations for certain γ values in the grid can be much higher; in

practice analysts should take advantage of the flexibility of the function to increase the number of iterations to ensure convergence. Generally, however, the method offers a useful framework for discovery of important variables in such settings.

A.2 Multistate Event Processes under Dependent Visit Scheme

To assess the sensitivity of the proposed method to violations of the conditional independence of visit process (CIVP) assumption ([Cook and Lawless, 2021](#)), we generate multistate data under a dependent visit scheme and then evaluate the variable selection performance of both the proposed multistate penalized EM algorithm and the penalized Cox regression model following imputation. The data is generated based on a joint model for the event process and the visit process, which is described in the following subsection.

Here we retain the notation $\{Z(t), 0 < t\}$ to represent the three-state progressive event process with state space $\{0, 1, 2\}$ and let $\{A(t), t < 0\}$ denote the visit process with state space $\{0, 1, 2, \dots\}$, where $A(t)$ counts the total number of followup visits made up to time t . We then define the joint process as $\{W(t) = Z(t)^{(A(t))}\}$ with state space $\{k^{(l)}, k = 0, 1, 2; l = 0, 1, 2, \dots\}$ to emphasize that $\{Z(t), 0 < t\}$ is the process of primary interest. A two-dimensional state space is given in [Figure A.1](#) for the joint model for a three-state progressive event process and the visit process. In this framework, horizontal movement from left to right in the diagram corresponds to the progression of the event process $\{Z(t), 0 < t\}$. However, progression in the event process is not detected until the subject makes a downward transition upon a visit. As in the CIVP setting the observed data is comprised of the times where there is an increment in $A(t)$ as well as $Z(t)$ at these

times.

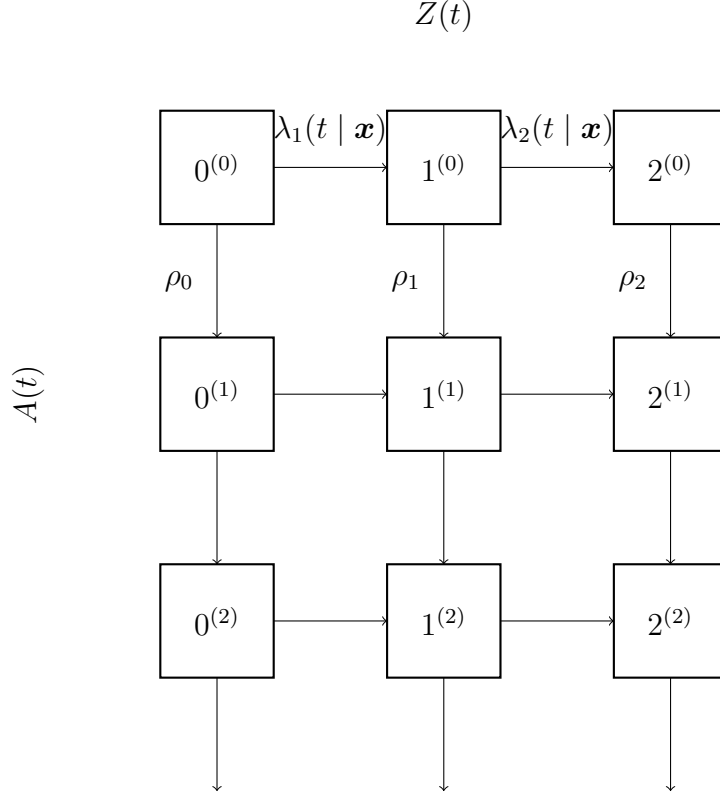


Figure A.1: A state space diagram for a joint model for a 3-state progressive event $Z(t)$ and a visit process $A(t)$.

The transition intensities for the event process are assumed to follow the same multiplicative form as specified in (2) and if $\mathcal{H}(t)$ is the history of the joint multistate and visit process, the visit process intensity is

$$\lim_{\Delta t \rightarrow 0} \frac{P(A(t + \Delta t^-) = l + 1 | A(t^-) = l, Z(t^-) = k, \mathcal{H}(t))}{\Delta t} = \rho_k(t) \quad (\text{A.1})$$

In what follows we adopt a simple model with $\rho_k(t) = \rho_k$ and set $\rho_2 = \rho_1 \exp(\gamma) = \rho_0 \exp(2\gamma)$, where γ characterizes the effect of state occupied on the visit process intensity.

When $\gamma > 0$, it indicates that the higher the state the subject reached in the event process, the more likely they are to make a visit. This scenario reflects situations where patients are more likely to visit the clinic as their disease worsens. When $\gamma < 0$, it suggests that subjects in higher states are less likely to make a visit. This could represent cases where patients are unable to visit the clinic as their disease becomes too severe. When $\gamma = 0$, the model reduces to the event process under an independent visit scheme, where the performance has been investigated and reported in the simulation studies section in the main text.

We considered $\gamma = \log 1.2$ and $\log 0.8$ under Settings 1, 2, and 3 for the regression coefficients β as described in Table 2.1. In order to focus on the effect of dependence between the event process and the visit process on variable selection performance and not conflate the effect of altering the overall frequency of visits, we solve for ρ_0 and λ_{01} such that $E(P(T_2 < \tau \mid \mathbf{X}; \theta)) = 0.95$ and $E(A(\tau)) = 4$ with $\tau = 1$ to control for the extent of right censoring and the expected number of visits during $(0, \tau]$. The average estimated true positive rates and false positive rates for analyzing multistate processes under the dependent visit scheme in Settings 1, 2, and 3 are summarized in Tables A.3, A.4, and A.5, respectively.

By comparing the results to those where the visit process is independent from the event process in Tables 2.1 and 2.2, we observed only slight changes in variable selection performance, regardless of whether γ is positive or negative. In general, when the CIVP assumption is violated, there is a mild degradation in variable selection performance reflected in lower true positive rates and higher false positive rates for the proposed penalized EM algorithm. For the Cox regression model, there is no consistent pattern whether the variable selection performance improves or deteriorates across settings and penalty functions when the CVIP assumption is violated, but the again changes are minimal. Overall,

these findings suggest that the proposed method is not highly sensitive to modest violations of the CIVP assumption given the same expected number of visits during the follow-up period.

					Penalized Cox Regression Following Imputation				
Multistate Analysis - Penalized EM Algorithm					Right-endpoint		Mid-point		
Metrics	LASSO	ALASSO	GLASSO	AGLASSO	Metrics	LASSO	ALASSO	LASSO	ALASSO
<i>Setting 1</i> *									
TP	96.6	93.8	99.5	97.9	—	—	—	—	—
TP ₁	98.5	97.5	99.5	97.9	TP ₁ ^{cox}	99.6	95.8	99.9	97.4
TP ₂	94.6	90.1	99.5	97.9	TP ₂ ^{cox}	99.6	95.8	99.9	97.4
TP _{cov}	100.0	99.7	99.5	97.9	TP _{cov} ^{cox}	99.6	95.8	99.9	97.4
FP	8.7	6.3	3.5	2.9	—	—	—	—	—
FP ₁	9.8	6.1	3.5	2.9	FP ₁ ^{cox}	9.1	10.5	8.4	10.1
FP ₂	7.5	6.5	3.5	2.9	FP ₂ ^{cox}	9.1	10.5	8.4	10.1
FP _{cov}	16.4	12.0	3.5	2.9	FP _{cov} ^{cox}	9.1	10.5	8.4	10.1
<i>Setting 2</i> †									
TP	97.1	95.3	98.1	95.6	—	—	—	—	—
TP ₁	99.6	99.8	100.0	99.2	TP ₁ ^{cox}	99.8	97.6	99.8	97.8
TP ₂	95.8	93.1	97.2	93.8	TP ₂ ^{cox}	89.0	77.8	89.1	79.3
TP _{cov}	98.7	97.7	97.2	93.8	TP _{cov} ^{cox}	89.0	77.8	89.1	79.3
FP	7.0	5.9	5.9	5.8	—	—	—	—	—
FP ₁	8.4	6.2	8.2	8.0	FP ₁ ^{cox}	11.8	12.7	12.2	12.4
FP ₂	5.5	5.6	3.4	3.5	FP ₂ ^{cox}	8.1	10.2	8.5	9.7
FP _{cov}	13.4	11.3	3.4	3.5	FP _{cov} ^{cox}	8.1	10.2	8.5	9.7
<i>Setting 3</i> ‡									
TP	61.5	64.9	97.2	94.4	—	—	—	—	—
TP ₁	97.8	97.7	97.2	94.4	TP ₁ ^{cox}	96.4	85.1	97.3	86.7
TP ₂	25.2	32.1	97.2	94.4	TP ₂ ^{cox}	96.4	85.1	97.3	86.7
TP _{cov}	97.9	98.2	97.2	94.4	TP _{cov} ^{cox}	96.4	85.1	97.3	86.7
FP	4.6	6.0	2.9	4.2	—	—	—	—	—
FP ₁	4.8	6.5	2.9	4.2	FP ₁ ^{cox}	7.2	8.7	7.3	9.6
FP ₂	4.5	5.6	2.9	4.2	FP ₂ ^{cox}	7.2	8.7	7.3	9.6
FP _{cov}	9.0	11.7	2.9	4.2	FP _{cov} ^{cox}	7.2	8.7	7.3	9.6

* Setting 1: $x_1, \dots, x_5$ and $x_{96}, \dots, x_{100}$ have strong effects of $\log(1.2)$ on both transitions, while other covariates are noise variables;

† Setting 2: $x_1, \dots, x_5$ have strong effects of $\log(1.2)$ on the $0 \rightarrow 1$ transition but no effects on the $1 \rightarrow 2$ transition; $x_{96}, \dots, x_{100}$ have strong effects of $\log(1.2)$ on both transitions, while other covariates are noise variables;

‡ Setting 3: $x_1, \dots, x_5$ and $x_{96}, \dots, x_{100}$ have strong effects of $\log(1.2)$ on $0 \rightarrow 1$ transition and weak effects of $\log(1.05)$ on $1 \rightarrow 2$ transition, while other covariates are noise variables.

Table A.1: Empirical results for multistate state analysis under independent intermittent observation scheme summarizing transition-specific and covariate-wise true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression following ad hoc right-endpoint or mid-point imputation; $n = 500$ with $\rho = 16$ as the mean number of visits over $(0, 1]$.

					Penalized Cox Regression Following Imputation				
Multistate Analysis - Penalized EM Algorithm					Right-endpoint		Mid-point		
Metrics	LASSO	ALASSO	GLASSO	AGLASSO	Metrics	LASSO	ALASSO	LASSO	ALASSO
<i>Setting 1</i> *									
TP	84.5	76.8	98.6	89.6	—	—	—	—	—
TP ₁	97.5	93.5	98.6	89.6	TP ₁ ^{cox}	89.4	71.4	97.0	84.0
TP ₂	71.4	60.1	98.6	89.6	TP ₂ ^{cox}	89.4	71.4	97.0	84.0
TP _{cov}	99.3	96.6	98.6	89.6	TP _{cov} ^{cox}	89.4	71.4	97.0	84.0
FP	6.9	4.7	3.7	2.2	—	—	—	—	—
FP ₁	8.9	4.0	3.7	2.2	FP ₁ ^{cox}	7.1	13.4	7.6	10.5
FP ₂	4.9	5.3	3.7	2.2	FP ₂ ^{cox}	7.1	13.4	7.6	10.5
FP _{cov}	13.3	8.9	3.7	2.2	FP _{cov} ^{cox}	7.1	13.4	7.6	10.5
<i>Setting 2</i> †									
TP	86.5	82.2	90.5	85.9	—	—	—	—	—
TP ₁	97.8	96.4	98.4	96.6	TP ₁ ^{cox}	94.8	80.0	98.2	90.2
TP ₂	80.9	75.1	86.5	80.6	TP ₂ ^{cox}	72.9	59.9	83.1	69.3
TP _{cov}	92.8	88.6	86.5	80.6	TP _{cov} ^{cox}	72.9	59.9	83.1	69.3
FP	6.8	6.2	4.4	4.4	—	—	—	—	—
FP ₁	8.9	6.7	6.3	6.0	FP ₁ ^{cox}	7.9	12.7	9.6	11.3
FP ₂	4.6	5.8	2.5	2.8	FP ₂ ^{cox}	5.5	11.2	6.4	9.3
FP _{cov}	12.8	11.5	2.5	2.8	FP _{cov} ^{cox}	5.5	11.2	6.4	9.3
<i>Setting 3</i> ‡									
TP	54.0	58.2	54.0	84.8	—	—	—	—	—
TP ₁	95.2	93.4	95.2	84.8	TP ₁ ^{cox}	79.0	60.7	91.3	73.5
TP ₂	12.7	23.0	12.7	84.8	TP ₂ ^{cox}	79.0	60.7	91.3	73.5
TP _{cov}	95.5	94.6	95.5	84.8	TP _{cov} ^{cox}	79.0	60.7	91.3	73.5
FP	5.0	7.1	5.0	2.3	—	—	—	—	—
FP ₁	6.3	7.9	6.3	2.3	FP ₁ ^{cox}	7.0	10.5	6.8	11.0
FP ₂	3.7	6.4	3.7	2.3	FP ₂ ^{cox}	7.0	10.5	6.8	11.0
FP _{cov}	9.7	13.7	9.7	2.3	FP _{cov} ^{cox}	7.0	10.5	6.8	11.0

* Setting 1: $x_1, \dots, x_5$ and $x_{96}, \dots, x_{100}$ have strong effects of $\log(1.2)$ on both transitions, while other covariates are noise variables;

† Setting 2: $x_1, \dots, x_5$ have strong effects of $\log(1.2)$ on the $0 \rightarrow 1$ transition but no effects on the $1 \rightarrow 2$ transition; $x_{96}, \dots, x_{100}$ have strong effects of $\log(1.2)$ on both transitions, while other covariates are noise variables;

‡ Setting 3: $x_1, \dots, x_5$ and $x_{96}, \dots, x_{100}$ have strong effects of $\log(1.2)$ on $0 \rightarrow 1$ transition and weak effects of $\log(1.05)$ on $1 \rightarrow 2$ transition, while other covariates are noise variables.

Table A.2: Empirical results for multistate state analysis under independent intermittent observation scheme summarizing transition-specific and covariate-wise true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression following ad hoc right-endpoint or mid-point imputation; $n = 500$ with $\rho = 4$ as the mean number of visits over $(0, 1]$.

						Penalized Cox Regression Following Imputation			
Metrics	Multistate Analysis - Penalized EM algorithm				Metrics	Right-endpoint		Mid-point	
	LASSO	ALASSO	GLASSO	AGLASSO		LASSO	ALASSO	LASSO	ALASSO
$\gamma = \log 1.2$									
TP	94.2	86.6	100.0	98.3	—	—	—	—	—
TP ₁	99.6	98.9	100.0	98.3	TP ₁ ^{cox}	97.6	86.6	99.7	95.4
TP ₂	88.7	74.3	100.0	98.3	TP ₂ ^{cox}	97.6	86.6	99.7	95.4
TP _{cov}	100.0	99.6	100.0	98.3	TP _{cov} ^{cox}	97.6	86.6	99.7	95.4
FP	7.1	2.3	2.2	0.9	—	—	—	—	—
FP ₁	9.4	1.4	2.2	0.9	FP ₁ ^{cox}	5.5	6.4	4.5	5.4
FP ₂	4.9	3.1	2.2	0.9	FP ₂ ^{cox}	5.5	6.4	4.5	5.4
FP _{cov}	13.8	4.5	2.2	0.9	FP _{cov} ^{cox}	5.5	6.4	4.5	5.4
$\gamma = \log 0.8$									
TP	95.2	86.4	100.0	97.5	—	—	—	—	—
TP ₁	100.0	98.8	100.0	97.5	TP ₁ ^{cox}	97.2	83.5	99.7	95.7
TP ₂	90.4	74.0	100.0	97.5	TP ₂ ^{cox}	97.2	83.5	99.7	95.7
TP _{cov}	100.0	99.6	100.0	97.5	TP _{cov} ^{cox}	97.2	83.5	99.7	95.7
FP	7.4	2.4	2.9	0.9	—	—	—	—	—
FP ₁	9.4	1.2	2.9	0.9	FP ₁ ^{cox}	4.9	7.2	4.7	5.5
FP ₂	5.4	3.6	2.9	0.9	FP ₂ ^{cox}	4.9	7.2	4.7	5.5
FP _{cov}	14.1	4.8	2.9	0.9	FP _{cov} ^{cox}	4.9	7.2	4.7	5.5

Table A.3: Empirical results for multistate state analysis under dependent intermittent observation scheme summarizing transition-specific and covariate-wise true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression following ad hoc right-endpoint or mid-point imputation; $n = 1000$ with $\rho = 4$ as the mean number of visits over $(0, 1]$ under Setting 1.

						Penalized Cox Regression Following Imputation			
Multistate Analysis - Penalized EM algorithm						Right-endpoint		Mid-point	
Metrics	LASSO	ALASSO	GLASSO	AGLASSO	Metrics	LASSO	ALASSO	LASSO	ALASSO
$\gamma = \log 1.2$									
TP	94.7	89.9	97.6	91.7	—	—	—	—	—
TP ₁	99.8	99.4	100.0	99.0	TP ₁ ^{cox}	99.6	91.8	100.0	98.6
TP ₂	92.2	85.2	96.4	88.0	TP ₂ ^{cox}	84.7	70.4	92.0	79.9
TP _{cov}	97.6	92.6	96.4	88.0	TP _{cov} ^{cox}	84.7	70.4	92.0	79.9
FP	4.8	2.3	4.9	2.8	—	—	—	—	—
FP ₁	6.4	2.1	7.2	4.8	FP ₁ ^{cox}	8.4	8.6	8.7	7.7
FP ₂	3.1	2.5	2.4	0.8	FP ₂ ^{cox}	5.0	6.4	4.5	4.7
FP _{cov}	9.1	4.2	2.4	0.8	FP _{cov} ^{cox}	5.0	6.4	4.5	4.7
$\gamma = \log 0.8$									
TP	94.9	91.8	98.1	93.0	—	—	—	—	—
TP ₁	100.0	99.8	100.0	99.8	TP ₁ ^{cox}	98.2	92.0	100.0	98.8
TP ₂	92.3	87.8	97.2	89.6	TP ₂ ^{cox}	82.2	68.3	91.4	78.1
TP _{cov}	97.7	95.5	97.2	89.6	TP _{cov} ^{cox}	82.2	68.3	91.4	78.1
FP	4.5	2.6	5.0	3.0	—	—	—	—	—
FP ₁	5.8	2.3	7.4	5.0	FP ₁ ^{cox}	7.0	8.1	8.5	8.1
FP ₂	3.2	2.9	2.5	0.9	FP ₂ ^{cox}	3.8	6.1	4.4	5.3
FP _{cov}	8.5	4.9	2.5	0.9	FP _{cov} ^{cox}	3.8	6.1	4.4	5.3

Table A.4: Empirical results for multistate state analysis under dependent intermittent observation scheme summarizing transition-specific and covariate-wise true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression following ad hoc right-endpoint or mid-point imputation; $n = 1000$ with $\rho = 4$ as the mean number of visits over $(0, 1]$ under Setting 2.

						Penalized Cox Regression Following Imputation			
Multistate Analysis - Penalized EM algorithm						Right-endpoint		Mid-point	
Metrics	LASSO	ALASSO	GLASSO	AGLASSO	Metrics	LASSO	ALASSO	LASSO	ALASSO
$\gamma = \log 1.2$									
TP	57.7	61.3	99.2	95.4	—	—	—	—	—
TP ₁	99.6	98.0	99.2	95.4	TP ₁ ^{cox}	92.6	76.3	98.6	88.2
TP ₂	15.7	24.6	99.2	95.4	TP ₂ ^{cox}	92.6	76.3	98.6	88.2
TP _{cov}	99.6	98.2	99.2	95.4	TP _{cov} ^{cox}	92.6	76.3	98.6	88.2
FP	2.8	2.2	2.2	0.8	—	—	—	—	—
FP ₁	3.6	1.9	2.2	0.8	FP ₁ ^{cox}	4.0	5.7	4.4	4.3
FP ₂	2.0	2.4	2.2	0.8	FP ₂ ^{cox}	4.0	5.7	4.4	4.3
FP _{cov}	5.5	4.3	2.2	0.8	FP _{cov} ^{cox}	4.0	5.7	4.4	4.3
$\gamma = \log 0.8$									
TP	57.5	58.3	99.8	94.9	—	—	—	—	—
TP ₁	99.9	99.1	99.8	94.9	TP ₁ ^{cox}	93.6	73.4	99.3	90.0
TP ₂	15.0	17.6	99.8	94.9	TP ₂ ^{cox}	93.6	73.4	99.3	90.0
TP _{cov}	99.9	99.2	99.8	94.9	TP _{cov} ^{cox}	93.6	73.4	99.3	90.0
FP	2.4	2.4	2.4	1.1	—	—	—	—	—
FP ₁	2.7	1.6	2.4	1.1	FP ₁ ^{cox}	4.2	5.5	4.7	5.2
FP ₂	2.2	3.2	2.4	1.1	FP ₂ ^{cox}	4.2	5.5	4.7	5.2
FP _{cov}	4.8	4.8	2.4	1.1	FP _{cov} ^{cox}	4.2	5.5	4.7	5.2

Table A.5: Empirical results for multistate state analysis under dependent intermittent observation scheme summarizing transition-specific and covariate-wise true positive and false positive rates (%) using the multistate penalized EM analysis and Cox regression following ad hoc right-endpoint or mid-point imputation; $n = 1000$ with $\rho = 4$ as the mean number of visits over $(0, 1]$ under Setting 3.

Appendix B

Appendix to Chapter 3

B.1 Computation of Conditional Probabilities

In this section, we show how to calculate the conditional expectations involved in the E-step of the EM algorithm. The subscript i denoting subjects and the dependence on $\boldsymbol{\theta}^{(r)}$ are suppressed for more concise expressions. The sufficient statistics we need in the complete data log-likelihood include N_{jkh} and u_{jkh} where $k \in S_j$, $j = 0, 1$. In order to evaluate the Q functions in (3.14) and (3.16), we need to derive expressions for the calculation of $\eta_{jkh} = E(N_{jkh} \mid D; \boldsymbol{\theta})$ and $\mu_{jkh} = E(u_{jkh} \mid D; \boldsymbol{\theta})$. We define $\mathcal{A}_r = [a_{r-1}, a_r)$ as the r th interval partitioned by the sequence of visit times for $r \leq m$, and $\mathcal{A}_{m+1} = [a_m, v)$. Let $\mathcal{C}_{khr} = \mathcal{A}_r \cap \mathcal{B}_{kh}$ denote the sub-interval of $\mathcal{A}_r$ within which the intensity for a transition into state k is constant, with $\mathcal{A}_r = \bigcup_{h=1}^{q_k} \mathcal{C}_{khr}$ for $k = 1, 2$, and $r = 1, 2, \dots, m$. Note also that and $\mathcal{B}_r = \bigcup_{r=1}^{m+1} \mathcal{C}_{khr}$. The expected number of $j \rightarrow k$ transitions within $\mathcal{B}_{kh}$ can be decomposed and written as the sum of expected transitions within $\mathcal{C}_{khr}$ over all intervals

between visits so

$$\eta_{jkh} = \sum_{r=1}^{m+1} \eta_{jkhr} \quad (\text{B.1})$$

where $\eta_{jkhr} = E(N_{jkhr} \mid D)$. Suppressing the dependence on covariates, $E(N_{jkhr} \mid D)$ is

$$P(Z(L_{khr}) = j, Z(R_{khr}^-) = k \mid D) = \frac{P(Z(a_1), \dots, Z(a_{r-1}), Z(L_{khr}), Z(R_{khr}), Z(a_r), \dots, Z(a_m), v, \delta)}{P(Z(a_1), \dots, Z(a_m), v, \delta)} \quad (\text{B.2})$$

which is given by

$$\frac{P(Z(L_{khr}) = j \mid Z(a_{r-1}))P(Z(R_{khr}) = k \mid Z(L_{khr}) = j)P(Z(a_r) \mid Z(R_{khr}^-) = k)}{P(Z(a_r) \mid Z(a_{r-1}))}$$

if $k = 1$ and $r \leq m$, and

$$\frac{P(Z(L_{khr}) = j \mid Z(a_m))P(Z(R_{khr}^-) = k \mid Z(L_{khr}) = j)P(Z(v^-) = j \mid Z(R_{khr}^-) = k)\lambda_{jk}(v)^\delta}{\sum_{l=1}^2 P(Z(v^-) = l \mid Z(a_m))\lambda_{lk}(v)^\delta}$$

if $k = 2$ or $r = m + 1$.

The expected time at risk of making a $j \rightarrow k$ transition within $\mathcal{B}_{kh}$ can similarly be partitioned as the sum of expected times at risk within $\mathcal{C}_{khr}$ over all intervals between visits. If $u_{jkhr} = \int_{\mathcal{C}_{khr}} I(Z(s) = j)ds$ then $u_{jkh} = \sum_{r=1}^{m+1} \mu_{jkhr}$ with

$$\begin{aligned} u_{jkhr} &= \int_{\mathcal{C}_{khr}} P(Z(s) = j \mid D^\circ)ds \\ &= \sum_{(z_L, z_R)} \int_{\mathcal{C}_{khr}} P(Z(s) = j \mid D, Z(L_{khr}) = z_L, Z(R_{khr}^-) = z_R, D^\circ) \\ &\quad \times P(Z(L_{khr}) = z_L, Z(R_{khr}^-) = z_R \mid D^\circ)ds \end{aligned} \quad (\text{B.3})$$

Note that this can be written as

$$\begin{aligned}
&= \left\{ P(Z(L_{khr}) = j, Z(R_{khr}^-) = j \mid D^\circ) \int_{\mathcal{C}_{khr}} P(Z(s) = j \mid Z(L_{khr}) = j, Z(R_{khr}^-) = j, D^\circ) + \right. \\
&\quad P(Z(L_{khr}) = k, Z(R_{khr}^-) = k \mid D^\circ) \int_{\mathcal{C}_{khr}} P(Z(s) = j \mid Z(L_{khr}) = k, Z(R_{khr}^-) = k, D^\circ) + \\
&\quad \left. P(Z(L_{khr}) = j, Z(R_{khr}^-) = k \mid D^\circ) \int_{\mathcal{C}_{khr}} P(Z(s) = j \mid Z(L_{khr}) = j, Z(R_{khr}^-) = k, D^\circ) \right\},
\end{aligned}$$

so

$$\begin{aligned}
\mu_{jkhr} &= \left\{ P(Z(L_{khr}) = j, Z(R_{khr}^-) = j \mid D^\circ) (L_{khr} - R_{khr}) \right. \\
&\quad + P(Z(L_{khr}) = k, Z(R_{khr}^-) = k \mid D^\circ) \cdot 0 \\
&\quad \left. + P(Z(L_{khr}) = j, Z(R_{khr}^-) = k \mid D^\circ) \int_{\mathcal{C}_{khr}} P(Z(s) = j \mid Z(L_{khr}) = j, Z(R_{khr}^-) = k, D^\circ) \right\}
\end{aligned}$$

The probabilities $P(Z(L_{khr}) = j, Z(R_{khr}^-) = k \mid D^\circ)$ are easy to compute as done in (B.2). In addition, we need

$$\int_{\mathcal{C}_{khr}} P(Z(s) = j \mid D^\circ, Z(L_{khr}) = j, Z(R_{khr}^-) = k) \quad (\text{B.4})$$

only when $P(Z(L_{khr}) = j, Z(R_{khr}^-) = k \mid D^\circ)$ is positive. This is computed as

$$\int_{\mathcal{C}_{khr}} \frac{P(Z(a_1), \dots, Z(a_{r-1}), Z(L_{khr}), Z(s) = j, Z(R_{khr}^-), Z(a_r), \dots, Z(a_m), v, \delta)}{P(Z(a_1), \dots, Z(L_{khr}), Z(R_{khr}^-), \dots, Z(a_m), v, \delta)} \quad (\text{B.5})$$

Variable	Original variable name	Definition given by NACC	Original coding	Recoding
SEX.FEMALE	SEX	Subject's sex	1 = Male 2 = Female	Binary
RACE.NONWHITE	RACE	Subject's race	Allowable codes 1 = White 2 = Black or African American 3 = American Indian or Alaska Native 4 = Native Hawaiian or Other Pacific Islander 5 = Asian 50 = Other (specify) 99 = Unknown	Exclude 99; 1 = white, 2 = other; binary
NACCNE4S	NACCNE4S	Number of APOE e4 alleles	Allowable Codes 0 = No e4 allele 1 = 1 copy of e4 allele 2 = 2 copies of e4 allele 9 = Missing/unknown/not assessed 99 = Unknown	Exclude 9, 1 = present, 0 = absent; binary
EDUC	EDUC	Years of education	12 = high school or GRE 16 = bachelor's degree 18 = master's degree 20 = doctorate 99 = Unknown	Exclude 99; treat as continuous
DECCLCOG	DECCLCOG	Cognitive impairment (clinician judgment)	0 = No, 1 = Yes, -4 = Not available	Exclude -4; binary
DECCLBE	DECCLBE	Behavioral symptoms (clinician judgment)	0 = No, 1 = Yes, -4 = Not available	Exclude -4; binary
DECCLMOT	DECCLMOT	Motor symptoms (clinician judgment)	0 = No, 1 = Yes, -4 = Not available	Exclude -4; binary
MOCATOTS	MOCATOTS	MoCA Total Raw Score (0-30)	88, -4 = Not available	Exclude 88, -4; continuous
COGSTAT	COGSTAT	Cognitive status (clinician)	0, -4, 9 = Not available	Exclude 0, -4, 9; treat as continuous
HIPPATR	HIPPATR	Hippocampal atrophy	0 = No, 1 = Yes, 8 = Unknown, -4 = Not available	Exclude 0, 8, -4; binary
CDRLANG	CDRLANG	Language impairment level	0-3 severity, -4 = Not available	Exclude -4; treat as continuous
CRAFTVRS	CRAFTVRS	Craft Story 21 Recall (Immediate)	-4, 95-98 = Issues/refusal	Exclude -4, 95-98; treat as continuous
VEG	VEG	Vegetables — Total number of vegetables named in 60 seconds	Allowable codes 0-77 95 = Physical problem 96 = Cognitive/behavior problem 97 = Other problem 98 = Verbal refusal -4 = Not available: UDS form submitted did not collect data in this way, or a skip pattern precludes response to this question	Exclude -4, 95, 96, 97, 98; treat as continuous
ANIMALS	ANIMALS	Animals — Total number of animals named in 60 seconds	Allowable codes 0-77 95 = Physical problem 96 = Cognitive/behavior problem 97 = Other problem 98 = Verbal refusal -4 = Not available: UDS form submitted did not collect data in this way, or a skip pattern precludes response to this question	Exclude -4, 95, 96, 97, 98; treat as continuous
TRAILA	TRAILA	Trail Making Test Part A — Total number of seconds to complete	Allowable codes 0 - 150 995 = Physical problem 996 = Cognitive/behavior problem 997 = Other problem 998 = Verbal refusal -4 = Not available: UDS form submitted did not collect data in this way, or a skip pattern precludes response to this question	Exclude -4, 995, 996, 997, 998; treat as continuous
CRAFTDTI	CRAFTDTI	Craft Story 21 Recall (Delayed) — Delay time	Allowable codes 0-85 99 = Unknown -4 = Not available: UDS form submitted did not collect data in this way, or a skip pattern precludes response to this question	Exclude -4, 99; continuous
MOCAREGI	MOCAREGI	MoCA: Memory — Registration (two trials)	Allowable codes 0 - 10 95 = Physical problem 96 = Cognitive/behavior problem 97 = Other problem 98 = Verbal refusal -4 = Not available: UDS form submitted did not collect data in this way, or a skip pattern precludes response to this question	Exclude 95, 96, 97, 98, -4; continuous
MINTTOTS	MINTTOTS	Multilingual Naming Test (MINT) — Total score	Allowable codes 0-32 95 = Physical problem 96 = Cognitive/behavior problem 97 = Other problem 98 = Verbal refusal -4 = Not available: UDS form submitted did not collect data in this way, or a skip pattern precludes response to this question	Exclude 95, 96, 97, 98, -4; continuous
COGMODE	COGMODE	Mode of onset of cognitive symptoms	Allowable codes 0 = No impairment in cognition 1 = Gradual 2 = Subacute 3 = Abrupt 4 = Other (specify) 99 = Unknown	Exclude 99, 4; treat as continuous
BEMODE	BEMODE	Mode of onset of behavioral symptoms	Allowable codes 0 = No behavioral symptoms 1 = Gradual 2 = Subacute 3 = Abrupt 4 = Other (specify) 99 = Unknown	Exclude 99, 4; treat as continuous

Table B.1: Description and preprocessing of selected candidate variables

B.2 Data Dictionary for Candidate Variables

B.3 Variable Selection Performance for Right-censored Data

In this section, we examine how various aspects of the selection algorithm affect performance under right-censored data conditions, where only the death time is right-censored and there are no intermittent observations. Specifically, we investigated two main aspects including (1) use of nonparametric and a weakly-parametric piecewise-constant baseline intensity, and (2) use of different penalty functions as well as ways to select the optimal tuning parameter(s).

This investigation complements our interval-censored data analysis in the main section in several ways. First, it allows us to explore additional modeling options and nuanced factors that could influence performance or explain differences between the proposed 3-state model and reference methods, which helps ensure fair comparisons and valid conclusions when we analyze the interval-censored multistate data. These options are not directly applicable to interval-censored cases but are straightforward to implement with right-censored data. Additionally, when the 0-1 process is under intermittent observation but as the number of assessments approaches infinity, the scenario converges to right-censored data where the entry time to state 1 is exactly observed and only the death time remains right-censored. Examining right-censored cases allows us to verify both our proposed algorithm and relative performance compared to the reference methods.

We first investigated whether different methods of estimating baseline intensities would significantly impact the estimates in right-censored data. We considered both the 3-state

illness-death model and failure time analysis for overall survival and progression-free survival, where the baseline estimation method could be nonparametric or piecewise-constant. For all failure time analysis, we used BIC where the likelihood is the partial likelihood, and n is the number of events which is the number of observed deaths for OS analysis and transitions to state 1 or death for PFS analysis.

For all methods with a piecewise-constant baseline intensity assumption, the cutpoints are selected such that there are roughly equal numbers of observations in each interval partitioned by cut-points. When comparing non-parametric methods against piecewise constant baseline estimation, we observe that the non-parametric approach generally outperforms the piecewise constant baseline due to its greater robustness to model misspecification. However, for the 3-state model, the piecewise constant approach performed better than the non-parametric approach because the non-parametric approach is not as efficient as the weakly-parametric piecewise constant approach when the model is correctly specified. For any of the three methods (3-state model, OS analysis, PFS analysis), the difference between non-parametric and piecewise constant approaches is minor.

We also explore different penalization approaches including a single LASSO, two separate LASSO penalties where each of them applies to each of the 0-1 process and the death process, and LASSO + GLASSO penalties where LASSO applies to the 0-1 process and GLASSO applies to the death process. Additionally, we also considering different methods for selecting optimal tuning parameters. Specifically, when there are two separate penalty functions, we can select the best model in two ways: One approach involves a two-dimensional grid search based on a BIC value based on the whole process. The alternative approach selects optimal tuning parameter values separately for each of the 0-1 process and the death process, resulting in two one-dimensional search for the optimal tuning parameters. These two approaches for selecting the best tuning parameter were

Method	Metric	True Positive (%)		False Positive (%)		
		non-parametric	piecewise-constant	non-parametric	piecewise-constant	
SETTING 1						
Penalized Multistate	TP_{01}	97.90	98.50	FP_{01}	1.27	1.17
	TP_{02}	95.60	98.20	FP_{02}	1.22	1.29
	TP_{12}	96.90	98.90	FP_{12}	4.70	3.06
	TP_{cov}	98.60	99.13	FP_{cov}	6.92	5.34
	TP_{coef}	97.04	98.60	FP_{coef}	2.37	1.83
	TP_2	98.80	99.60	FP_2	5.87	4.32
	TP_0	98.70	98.90	FP_0	2.48	2.41
Cox-OS	TP_2^{cox}	83.90	81.20	FP_2^{cox}	1.27	0.97
	TP^{cox}	68.10	66.45	FP^{cox}	1.14	0.88
Cox-PFS	TP_0^{cox}	95.00	94.80	FP_0^{cox}	0.93	0.88
SETTING 2						
Penalized Multistate	TP_{01}	96.60	97.80	FP_{01}	0.57	0.67
	TP_{02}	95.00	96.60	FP_{02}	0.38	0.77
	TP_{12}	97.00	98.80	FP_{12}	2.47	1.83
	TP_{cov}	96.20	97.80	FP_{cov}	3.48	3.28
	TP_{coef}	96.20	97.73	FP_{coef}	1.14	1.09
	TP_2	96.00	97.80	FP_2	2.84	2.64
	TP_0	95.80	97.20	FP_0	0.99	1.46
Cox-OS	TP_2^{cox}	75.50	68.70	FP_2^{cox}	0.91	0.51
	TP^{cox}	51.60	46.40	FP^{cox}	0.74	0.44
Cox-PFS	TP_0^{cox}	85.10	83.00	FP_0^{cox}	0.86	0.57
SETTING 3						
Penalized Multistate	TP_{01}	98.00	98.40	FP_{01}	0.98	1.09
	TP_{02}	96.80	97.20	FP_{02}	1.20	1.72
	TP_{12}	97.60	99.40	FP_{12}	4.57	4.36
	TP_{cov}	98.07	98.80	FP_{cov}	6.47	6.93
	TP_{coef}	97.60	98.56	FP_{coef}	2.23	2.37
	TP_2	98.40	99.40	FP_2	5.72	6.04
	TP_0	98.50	98.70	FP_0	2.18	2.71
Cox-OS	TP_2^{cox}	14.60	7.30	FP_2^{cox}	0.20	0.04
	TP^{cox}	7.55	3.70	FP^{cox}	0.19	0.05
Cox-PFS	TP_0^{cox}	24.53	24.00	FP_0^{cox}	0.22	0.18

Table B.2: True and false positive rates (%) for all methods across three settings. Penalized multistate results are reported with both non-parametric and piecewise-constant baselines; Cox-based models include OS and PFS settings.

Method	LASSO	True Positive (%)				LASSO	False Positive (%)				
		LASSO ¹		LASSO+GLASSO ²			LASSO ¹		LASSO+GLASSO ²		
		one-dim ³	two-dim ⁴	one-dim ³	two-dim ⁴		one-dim ³	two-dim ⁴	one-dim ³	two-dim ⁴	
SETTING 1											
TP_{01}	98.50	98.80	98.70	98.80	98.70	FP_{01}	1.17	1.80	1.77	1.80	1.77
TP_{02}	98.20	98.40	97.60	99.40	99.20	FP_{02}	1.29	0.73	0.97	5.67	5.29
TP_{12}	98.90	97.90	95.90	97.90	96.10	FP_{12}	3.06	1.92	1.74	0.63	0.42
TP_{cov}	99.13	98.87	98.07	98.33	97.20	FP_{cov}	5.34	4.33	4.41	2.36	2.16
TP_{coef}	98.60	98.36	97.36	98.56	97.76	FP_{coef}	1.83	1.47	1.48	2.76	2.55
TP_2	99.60	99.00	97.80	97.90	96.10	FP_2	4.32	2.63	2.72	0.63	0.42
TP_0	98.90	99.20	99.00	99.20	99.10	FP_0	2.41	2.51	2.71	7.60	7.23
SETTING 2											
TP_{01}	97.80	97.80	97.80	97.80	97.80	FP_{01}	0.67	0.73	0.72	0.73	0.72
TP_{02}	96.60	95.80	97.00	92.80	93.00	FP_{02}	0.77	0.62	0.82	5.68	5.54
TP_{12}	98.80	98.40	96.60	96.20	94.00	FP_{12}	1.83	1.37	1.28	5.51	5.48
TP_{cov}	97.80	97.40	97.20	95.60	94.93	FP_{cov}	3.28	2.68	2.86	1.31	1.35
TP_{coef}	97.73	97.33	97.13	95.60	94.93	FP_{coef}	1.09	0.91	0.94	3.97	3.91
TP_2	97.80	97.20	96.80	94.50	93.50	FP_2	2.64	2.02	2.13	0.66	0.62
TP_0	97.20	96.80	97.50	95.30	95.40	FP_0	1.46	1.31	1.56	6.58	6.50
SETTING 3											
TP_{01}	98.40	98.80	98.50	98.80	98.50	FP_{01}	1.09	1.70	2.07	1.70	2.07
TP_{02}	97.20	95.60	96.60	100.00	100.00	FP_{02}	1.72	1.03	1.03	5.61	5.62
TP_{12}	99.40	99.00	97.40	97.80	97.00	FP_{12}	4.36	2.76	2.31	0.61	0.71
TP_{cov}	98.80	98.87	97.73	98.00	97.20	FP_{cov}	6.93	5.35	5.32	2.25	2.75
TP_{coef}	98.56	98.24	97.68	98.64	98.20	FP_{coef}	2.37	1.81	1.79	2.69	2.85
TP_2	99.40	99.20	97.70	97.80	97.00	FP_2	6.04	3.82	3.34	0.61	0.71
TP_0	98.70	99.10	98.90	99.20	98.80	FP_0	2.71	2.66	3.06	7.43	7.82

¹ An alternative LASSO with two tuning parameters, each for the $0 \rightarrow 1$ transitions and the death process, respectively.

² LASSO for the $0 \rightarrow 1$ process and GLASSO for the $0 \rightarrow 2$ and $1 \rightarrow 2$ process.

³ Two separate one-dimensional grid searches with one for the tuning parameter for $0 \rightarrow 1$ process and the other for the tuning parameter for transitions to state 2.

⁴ A two-dimensional grid search for the best combination of the two tuning parameters.

Table B.3: True and false positive rates (%) for the LASSO-based models across three settings.

also considered for the LASSO+GLASSO penalty.

It's worth noting that implementing two one-dimensional searches with interval-censored data under the mixed observation scheme is not straightforward. When the process is subject only to right-censoring, the observed likelihood can be factored into two parts, where one corresponds to the 0-1 process and the other to the death process, allowing for separate hyperparameter tuning. We observed that generally, the two one-dimensional grid search performed slightly better than the single two-dimensional grid search.

Appendix C

Appendix to Chapter 4

C.1 Derivations for Expectations Involved for Calculating Limiting Values for Cox Regression.

Here we derive the closed forms for the expectations involved in the estimating equation for the Cox regression (4.12). In what follows, we assume that the Cox regression adjusts for a covariate vector X_1 , while the true underlying data-generating process depends on a covariate vector X with $X_1 \subseteq X$. This encompasses both scenarios we consider: (1) when the model includes all relevant covariates (i.e., $X_1 = X$), and (2) when important covariates are omitted from the model (i.e., $X_1 \subset X$). The following derivations hold for any X_1 and X with $X_1 \subseteq X$, hence we adopt generic notation by treating X_1 as a vector. We discuss the derivations under two main settings, including (1) when the true process is a three-state Markov process, and (2) when the true process is a three-state semi-Markov process.

C.1.1 Cox Regression in a Three-state Markov Process

Given the censoring time is independent of the event process conditional on X , we have

$$\begin{aligned}
r^{(0)}(s) &= E [\bar{Y}(s) dN_2(s)] \\
&= E [Y^\dagger(s) Y(s) dN_2(s)] \\
&= E_X \left\{ E_{Y^\dagger(s)} \left\{ E_{dN_2(s)} [E_{Y(s)} [Y^\dagger(s) Y(s) dN_2(s) \mid Y^\dagger(s), dN_2(s), X] \mid Y^\dagger(s), X] \mid X \right\} \right\} \\
&= E_X \left\{ G(s \mid X) E_{Y^\dagger(s)} \left\{ E_{dN_2(s)} [Y^\dagger(s) dN_2(s) \mid Y^\dagger(s), X] \mid X \right\} \right\},
\end{aligned}$$

where $G(s \mid X) = P(C > s \mid X)$. The innermost expectation is

$$E_{dN_2(s)} [Y^\dagger(s) dN_2(s) \mid Y^\dagger(s), X] = Y^\dagger(s) \cdot E_{dN_2(s)} [dN_2(s) \mid Y^\dagger(s), X], \quad (\text{C.1})$$

which is equal to

$$\begin{aligned}
Y^\dagger(s) \Big\{ & E_{dN_2(s)} [dN_2(s) \mid Z(s^-) = 0, X] P(Z(s^-) = 0 \mid Z(s^-) < 2, X) \\
& + E_{dN_2(s)} [dN_2(s) \mid Z(s^-) = 1, X] P(Z(s^-) = 1 \mid Z(s^-) < 2, X) \Big\}.
\end{aligned} \quad (\text{C.2})$$

Since the true process is Markov, the conditional expectation $E_{dN_2(s)} [dN_2(s) | Z(s^-) = j, X]$ is simply $\lambda_{j2}(s | X) ds$ for $j = 0, 1$. Therefore, we have

$$\begin{aligned}
r^{(0)}(s) &= E_X \left\{ G(s | X) E_{Y^\dagger(s)} \left[Y^\dagger(s) \sum_{j=0}^1 \lambda_{j2}(s | X) ds \cdot P(Z(s^-) = j | Z(s^-) < 2, X) \right] \right\} \\
&= E_X \left\{ G(s | X) P(Z(s^-) < 2 | X) \sum_{j=0}^1 \lambda_{j2}(s | X) ds \cdot P(Z(s^-) = j | Z(s^-) < 2, X) \right\} \\
&= E_X \left\{ G(s | X) \sum_{j=0}^1 \lambda_{j2}(s | X) ds \cdot P(Z(s^-) = j | X) \right\} \\
&= E_X \left\{ G(s | X) \sum_{j=0}^1 \lambda_{j2}(s | X) ds \cdot P(Z(s^-) = j | X) \right\}.
\end{aligned}$$

Similarly, we can rewrite $r^{(1)}(s)$ as

$$\begin{aligned}
r^{(1)}(s) &= E [\bar{Y}(s) X_1 dN_2(s)] \\
&= E_X \left[G(s | X) \sum_{j=0}^1 P(Z(s^-) = j | X) X_1 \lambda_{j2}(s | X) ds \right], \tag{C.3}
\end{aligned}$$

$$\begin{aligned}
r^{(0)}(s; \psi) &= E [\bar{Y}(s) \exp(X'_1 \psi)] \\
&= E_X [G(s | X) P(Z(s^-) < 2 | X) \exp(X'_1 \psi)], \tag{C.4}
\end{aligned}$$

and

$$\begin{aligned}
r^{(1)}(s; \psi) &= E [\bar{Y}(s) X_1 \exp(X'_1 \psi)] \\
&= E_X [G(s | X) X_1 \exp(X'_1 \psi) \cdot P(Z(s^-) < 2 | X)]. \tag{C.5}
\end{aligned}$$

The dimension of the estimating equation (4.12) depends on the dimension of X_1 . One can then evaluate the expectations given the censoring distribution, all parameters char-

acterizing the 3-state illness-death Markov process, and the distribution of the covariates X .

C.1.2 Cox Regression in a Semi-Markov Process

Here we consider a true process as a three-state semi-Markov process where the survival function for the sojourn time in state 1 is governed by the survival distribution

$$\mathcal{F}(w \mid X_1) = \exp(-(\lambda_{12}w)^{\kappa_{12}}) \exp(X'\beta_{12})$$

as described in Section 4.3.2, the conditional expectation $E_{dN_2(s)} [dN_2(s) \mid Z(s^-) = 1, X]$ depends not only on the current state $Z(s^-)$ and covariate X , but also on the sojourn time spent in the current state. Therefore, we condition on T_1 which allows us to determine the sojourn time at state 1 at time s as $s - T_1$, and write

$$\begin{aligned} & E_{dN_2(s)} [dN_2(s) \mid Z(s^-) = 1, X] \\ &= E_{T_1} \{ E_{dN_2(s)} [dN_2(s) \mid Z(s^-) = 1, T_1, X] \mid Z(s^-) = 1, X \}, \end{aligned}$$

where

$$\begin{aligned} & E_{dN_2(s)} [dN_2(s) \mid Z(s^-) = 1, T_1, X] \\ &= \lambda_{12}\kappa_{12} [\lambda_{12}(s - T_1)]^{\kappa_{12}-1} \exp(X'\beta_{12}). \end{aligned} \tag{C.6}$$

With this we can write the conditional expectation w.r.t. T_1 as

$$\begin{aligned} & E_{T_1} \{ \lambda_{12}\kappa_{12} \{ [\lambda_{12}(s - T_1)]^{\kappa_{12}-1} \} \exp(X'\beta_{12}) \mid Z(s^-) = 1, X \} \\ &= \lambda_{12}^{\kappa_{12}} \kappa_{12} \exp(X'\beta_{12}) \cdot E_{T_1} [(s - T_1)^{\kappa_{12}-1} \mid Z(s^-) = 1, X] \end{aligned}$$

which involves computing $E_{T_1} [(s - T_1)^{\kappa_{12}-1} | Z(s^-) = 1, X]$. Note that

$$\begin{aligned} & E_{T_1} [(s - T_1)^{\kappa_{12}-1} | Z(s^-) = 1, X] \\ &= \frac{1}{P(Z(s^-) = 1 | X)} \int_0^s (s - s_1)^{\kappa_{12}-1} \lambda_{01} \exp(X' \beta_{01}) \times \\ & \exp \{ - [\lambda_{01} \exp(X' \beta_{01}) + \lambda_{02} \exp(X' \beta_{02})] s_1 - \lambda_{12}^{\kappa_{12}} \exp(X' \beta_{12}) (s - s_1)^{\kappa_{12}} \} ds, \end{aligned}$$

so if we let

$$\begin{aligned} g(X) &= \int_0^s (s - s_1)^{\kappa_{12}-1} \lambda_{01} \exp(X' \beta_{01}) \\ & \times \exp \{ - [\lambda_{01} \exp(X' \beta_{01}) + \lambda_{02} \exp(X' \beta_{02})] s_1 - \lambda_{12}^{\kappa_{12}} \exp(X' \beta_{12}) (s - s_1)^{\kappa_{12}} \} ds, \end{aligned}$$

we have derived

$$E_{dN_2(s)} [dN_2(s) | Z(s^-) = 1, X] = g(X) / P(Z(s^-) = 1 | X) \quad (\text{C.7})$$

Plugging (C.7) back into (C.1) gives

$$r^{(0)}(s) = E_X [G(s | X) \lambda_{02}(s | X) ds P(Z(s^-) = 0 | X)] + E_X [G(s | X) g(X)], \quad (\text{C.8})$$

and

$$r^{(1)}(s) = E_X [G(s | X) \lambda_{02}(s | X) ds X_1 P(Z(s^-) = 0 | X)] + E_X [G(s | X) g(X) X_1]. \quad (\text{C.9})$$

Note that $r^{(0)}(s; \psi)$ and $r^{(1)}(s; \psi)$ remain the same as (C.4) and (C.5), respectively, as shown in the derivations for the scenarios where the true process is Markov.

C.2 Derivations for Calculating Limiting Value for 3-state Model

We present the derivations for the expectations involved in (4.10). Similarly, we assume that the 3-state model adjusts for covariate X_1 , while the true underlying data-generating process depends on a covariate vector X , with $X_1 \subseteq X$. We discuss the derivations under two main settings, including (1) when the true process is a three-state Markov process, and (2) when the true process is a three-state semi-Markov process.

C.2.1 3-state Markov Illness-death model in a Markov Process

Given that the censoring time is independent of the event process conditional on X ,

$$\begin{aligned}
 r_{jk}^{(0)}(s) &= E[\bar{Y}_j(s) dN_{jk}(s)] \\
 &= E[Y(s) Y_j(s) dN_{jk}(s)] \\
 &= E_X \left\{ E_{Y(s)} \left[E_{dN_{jk}(s)} \left\{ E_{Y_j(s)} [Y(s) Y_j(s) dN_{jk}(s) \mid Y_j(s), dN_{jk}(s), X] \mid Y_j(s), X \right\} \mid X \right] \right\} \\
 &= E_X \left\{ G(s \mid X) E_{Y_j(s)} \left[E_{dN_{jk}(s)} [Y_j(s) dN_{jk}(s) \mid Y_j(s), X] \mid X \right] \right\}
 \end{aligned} \tag{C.10}$$

for $j < k, j, k = 0, 1, 2$. Under the Markov assumption, the innermost conditional expectation can be written as

$$\begin{aligned}
 E_{dN_{jk}(s)} [Y_j(s) dN_{jk}(s) \mid Y_j(s), X] &= Y_j(s) E_{dN_{jk}(s)} [dN_{jk}(s) \mid Y_j(s), X] \\
 &= Y_j(s) E_{dN_{jk}(s)} [dN_{jk}(s) \mid Z(s^-) = j, X] \\
 &= Y_j(s) \lambda_{jk}(s \mid X) ds
 \end{aligned} \tag{C.11}$$

Therefore,

$$\begin{aligned}
r_{jk}^{(0)}(s) &= E_X \left\{ G(s | X) E_{Y_j^\dagger(s)} \left[Y_j^\dagger(s) \lambda_{jk}(s | X) ds \right] | X \right\} \\
&= E_X \left\{ G(s | X) P(Z(s^-) = j | X) \lambda_{jk}(s | X) ds \right\} \\
&= E_X \left\{ G(s | X) \lambda_{jk}(s | X) ds \cdot P(Z(s^-) = j | X) \right\}
\end{aligned} \tag{C.12}$$

Similarly,

$$r_{jk}^{(1)}(s) = E_X \left\{ G(s | X) \lambda_{jk}(s | X) ds X_1 P(Z(s^-) = j | X) \right\} \tag{C.13}$$

$$\begin{aligned}
r_{jk}^{(0)}(s; \psi) &= E[\bar{Y}_j(s) \exp(X_1' \psi)] \\
&= E_X \left\{ G(s | X) \exp(X_1' \psi) P(Z(s^-) = j | X) \right\}
\end{aligned} \tag{C.14}$$

$$\begin{aligned}
r_{jk}^{(1)}(s; \psi) &= E[\bar{Y}_j(s) X_1 \exp(X_1' \psi)] \\
&= E_X \left\{ G(s | X) X_1 \exp(X_1' \psi) P(Z(s^-) = j | X) \right\}.
\end{aligned} \tag{C.15}$$

C.2.2 Markov Illness-death Model in a Three-state Semi-Markov Process

When the true process is a three-state semi-Markov process where the survival function for the sojourn time in state 1 is set to follow a Weibull distribution as described in Section 4.3.2, the limiting value for ψ_{12} is affected because the conditional expectation $E_{dN_{12}(s)}[dN_{12}(s) | Y_1(s) = 1, X]$ now depends not only on the current state $Z(s^-)$ and co-

variate X , but also on the sojourn time spent in state 1. As shown in Section C.2, the limiting values ψ_{01} and ψ_{02} are not determined by the parameters governing the $1 \rightarrow 2$ process and thus remain unchanged from the Markov setting. Here we show how to derive ψ_{12}^* . We begin with equation (C.10) in Section C.2 and show that, under the semi-Markov assumption, the conditional expectation of $dN_{12}(s)$ can be expressed as

$$\begin{aligned} E_{dN_{12}(s)} [Y_1(s) dN_{12}(s) \mid Y_1(s), T_1, X] &= Y_1(s) E_{dN_{12}(s)} [dN_{12}(s) \mid Y_1(s), T_1, X] \\ &= Y_1(s) E_{dN_{12}(s)} [dN_{12}(s) \mid Z(s^-) = 1, T_1, X] \\ &= Y_1(s) \lambda_{12} \kappa_{12} [\lambda_{12}(s - T_1)]^{\kappa_{12}-1} \exp(X' \beta_{12}). \end{aligned}$$

Following the same derivations as described in Section C.1.2, we have

$$r^{(0)}(s) = E_X [G(s \mid X)g(X)], \quad (\text{C.16})$$

$$r_{jk}^{(1)}(s) = E_X [G(s \mid X)X_1], \quad (\text{C.17})$$

$$\begin{aligned} r_{jk}^{(0)}(s; \psi) &= E[\bar{Y}_1(s) \exp(X_1' \psi)] \\ &= E_X [G(s \mid X) \exp(X_1' \psi) P(Z(s^-) = 1 \mid X)], \end{aligned} \quad (\text{C.18})$$

and

$$\begin{aligned} r^{(1)}(s)(s; \psi) &= E[\bar{Y}_1(s)X_1 \exp(X_1' \psi)] \\ &= E_X [G(s \mid X) \exp(X_1' \psi)X_1 P(Z(s^-) = 1 \mid X)], \end{aligned} \quad (\text{C.19})$$

Computing the state occupancy probabilities under the semi-Markov model is described in Section C.2.3.

C.2.3 State Occupancy Probabilities in a Semi-Markov Process

Here we present the state occupancy probabilities used for computing limiting values when the true process is a 3-state Markov process as described in Section 4.3.2. The state 1 occupancy probability can be expressed as

$$\begin{aligned}
P(Z(t) = 1 \mid X) &= \int_0^t \exp \{ - [\Lambda_{01}(0, s_1 \mid X) + \Lambda_{02}(0, s_1 \mid X)] \lambda_{01}(s_1) \exp(-\Lambda_{12}(s_1, t \mid X)) \} ds_1 \\
&= \int_0^t \exp \left[-(\lambda_{01}e^{X'\beta_{01}} + \lambda_{02}e^{X'\beta_{02}})s_1 \right] \lambda_{01}e^{X'\beta_{01}} \exp \left[- \int_{s_1}^t \lambda_{12}(u \mid H(u), X) du \right] ds_1 \\
&= \int_0^t \exp \left[-(\lambda_{01}e^{X'\beta_{01}} + \lambda_{02}e^{X'\beta_{02}})s_1 \right] \lambda_{01}e^{X'\beta_{01}} \exp \left[- \int_{s_1}^t h_{12}(u - s_1)e^{X'\beta_{12}} du \right] ds_1 \\
&= \int_0^t \exp \left[-(\lambda_{01}e^{X'\beta_{01}} + \lambda_{02}e^{X'\beta_{02}})s_1 \right] \lambda_{01}e^{X'\beta_{01}} \exp \left[- \int_0^{t-s_1} h_{12}(w)e^{X'\beta_{12}} dw \right] ds_1 \\
&= \int_0^t \lambda_{01}e^{X'\beta_{01}} \exp \left[-(\lambda_{01}e^{X'\beta_{01}} + \lambda_{02}e^{X'\beta_{02}})s_1 - \lambda_{12}^{\kappa_{12}} e^{X'\beta_{12}}(t - s_1)^{\kappa_{12}} \right] ds_1.
\end{aligned}$$

The state 0 occupancy probability is

$$\begin{aligned}
P(Z(t) = 0 \mid X) &= \exp [-\Lambda_{01}(0, t \mid X) - \Lambda_{02}(0, t \mid X)] \\
&= \exp \left[-(\lambda_{01}e^{X'\beta_{01}} + \lambda_{02}e^{X'\beta_{02}})t \right],
\end{aligned}$$

and the state 2 occupancy probability can be calculated by

$$P(Z(t) = 2 \mid X) = 1 - P(Z(t) = 1 \mid X) - P(Z(t) = 0 \mid X).$$