Dynamic Prediction of Acute Graft-versus-Host-Disease with Longitudinal Biomarkers

by

Yumeng Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2018

Doctoral Committee:

Professor Thomas M. Braun, Chair
Professor Douglas E. Schaubel
Professor Kerby A. Shedden
Professor Jeremy M.G. Taylor

Yumeng Li

yumeng@umich.edu

ORCID id: 0000-0001-8898-1525

To Yumeng in 2013

you made an insane and righteous decision

which makes you proud of yourself for the rest of your life

ACKNOWLEDGEMENTS

My deep gratitude goes first to my advisor, Prof. Thomas Braun, who expertly guided me through my graduation education and who shared the excitement of five years of discovery. His guidance, encouragement, advice and patience kept me consistently engaged with my research, and his personal generosity and good humor helped make my PhD time enjoyable.

My appreciation also extends to my CSCAR colleagues. Many thanks to my dissertation committee member and GSRA supervisor, Prof. Kerby Shedden. His extensive knowledge and insights in statistics inspired me and helped me grow into an independent consultant. Thanks Prof. Brenda Gillespie, for supporting me during my PhD education and shared with me your research and life experience. My sincere gratitude also goes to my CSCAR colleagues, Kim Ward, Corey Powell, Josh Erickson, Michael Clark, Alex Cao, Hyungjin Myra Kim, Marcio Duarte Albasini Mourao. Thank you my peer GSRAs, Sophie Chen, Joseph Naiman, and Michael David Hornstein.

To my other committee members, Prof. Douglas Schaubel and Prof. Jeremy Taylor, thank you for your valuable questions and feedback, which have honed my thinking and improved the quality of this dissertation. I must also thank the great faculties, Prof. Michael Boehnke, Prof. Min Zhang. Prof. Lu Wang, Prof. Peter Song, Prof. Timothy Johnson, Prof. Yun Li, and Prof. Veronica Berrocal. Thank

you for sharing your valuable experience and knowledge in statistical research.

I would like to thank my fellows and friends in the Biostatistics Department, especially Jingchunzi Shi, Wenting Cheng, Xin Wang, Sheng Qiu, Paul Imbriano, Alan Wong, Yebin Tao, Ye Yang, Meng Xia, Boxian Wei, Tingting Lu, Rounak Dey, Cui Guo, Lu Tang, Xin Wang, Nicole Fenech and Daniel Muenz.

Thank you, my dear friend Fang Fang, for sharing all the up and downs with me. Thank you, Kirsten Herold, Yvonne Deporre and Benjamin Gebarski, for your effort in my writing.

Thanks also goes to the amusing souls which accompanied me in the hardest time. Thank you, William Somerset Maugham, Mary Roach, Richard Dawkins, Yuval Noah Harari, Edmond De Amicis, and Rick and Morty. Thank you for sharing your deepest understanding of life and human nature with me, so that I am looking at the stars no matter where I am.

Most of all, I am fully indebted to my parents, for their consistently support, encouragement, unconditional love, and 23 pairs of chromosomes. Their stubborn confidence in me nourished my research and let it blossom.

Thanks to all the people who made this dissertation possible, especially who is reading this. Yes, I mean you.

Thank you.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

This dissertation builds three prediction tools to dynamically predict the onset of acute graft-versus-host disease (aGVHD) with longitudinal biomarkers. Acute graft-versus-host disease is a complication for patients who have received allogeneic bone marrow transplant, and it is fatal for some patients. Clinicians could benefit from these prediction tools to identify patients who are at risk and who are not, and thus assign appropriate interventions.

Our first project introduces how to apply joint modeling with latent classes (JMLC) and landmark analysis to aGVHD data. In JMLC, we group all aGVHD-free patients into one latent class and define that class as the "cure" class. In landmark analysis, we incorporate patients' biomarker information up to the landmark time to gain more efficiency. Computer simulations show that both methods adjust for the measurement error, and that JMLC outperforms landmark analysis when the functional form of the biomarker profile is correctly specified.

In our second project, we describe how to execute dynamic prediction with the pattern mixture model, in which each patient is classified by his/her time-to-aGVHD, and patients in the same group share the same mean profile of biomarkers. The pattern mixture model is easy to execute and straightforward to interpret. Simulations indicate that the pattern mixture model controls loss of accuracy in predictions.

In our third project, we incorporate censored cases to generalize the pattern mix-

ture model in the second project. The simulation results demonstrate that this generalized pattern mixture model accurately estimates of the marginal pattern probabilities, and thus better estimates early predictions compared to early predictions not incorporating censored observations.

In our fourth project, we explain the process of parametric bootstrap in selecting the number of latent classes in JMLC. Compared with the standard information-based criteria in model selection in JMLC, our parametric bootstrap likelihood ratio test (LRT) controls the Type I error well while maintaining sufficient power. We also propose two sequential early stopping rules to relieve the computational burden of bootstrap.

CHAPTER I

Introduction

1.1 Acute Graft-versus-Host Disease

Approximately every three minutes one person is diagnosed with a hematologic cancer (blood cancer) in the United States (US), and approximately 160 people each day die from a hematologic cancer in the US (the Leukemia and Lymphoma Society, 2016). Based on Cancer Facts & Figures (2016) released by National Cancer Institute, over 60,000 Americans are expected to be diagnosed with leukemia, one of major hematologic cancers together with lymphomas and myeloma. There are multiple treatment strategies for hematologic cancer, such as chemotherapy, radiation therapy, immunotherapy, and hematologic stem-cell transplantation (HSCT). Among them, HSCT is a rapidly evolving technique that offers a potential cure to hematologic cancers and other hematologic disorders, such as primary immunodeficiency, aplastic anemia, and myelodysplasia.

There are two main types of HSCT, autologous HSCT and allogeneic HSCT. Many factors contribute to the choice of the two types of HSCT, including the type of cancer, the stage of cancer, the age of a patient, and the accessibility of matched donors (Champlin, 2003). In autologous HSCT, a patient's own stem cells are collected and frozen prior to the high-dose chemotherapy, and then are reinfused. There are rarely

graft failures and virtually no risk of graft-versus-host disease (GVHD), which is an inflammatory disease caused by immune cells in the donor's organ viewing the recipients' tissues as foreign and attacking them. Thus, the treatment-related mortality rates of autologous HSCT patients are low. However, these patients are at increased risk of relapse because there is a possibility that the graft is contaminated with tumor cells.

Patients undergoing allogeneic HSCT receive stem cells from human leukocyte antigen (HLA)-matched donors, who can be either siblings or unrelated donors. HLA is a cell-surface protein that regulates the human immune system. The primary benefit of allogeneic HSCT is that the graft is presumed to be tumor-free and there is no prior marrow injury from chemotherapy. Moreover, there is an additional graft-versus-tumor effect contributing to a lower recurrence rate. Hosing et al. (2003) found that the probability of non-Hodgkin lymphoma (NHL) recurrence is 19% among allogeneic HSCT patients, compared with 74% in autologous HSCT patients (p-value = 0.003). However, the overall survival after HSCT is not satisfactory (Center for International Blood & Marrow Transplant Research, 2015). For example, as shown in Figure 1.1, the three-year overall survival of patients after allogeneic HSCT with early-stage acute myelogenous leukemia (AML) is around 60%, while for advanced-stage patients it is only about 30%.

The main causes of mortality after allogeneic HSCT are relapse of cancer, GVHD, infections and other complications, as shown in Figure 1.2. GVHD is one of the major causes of non-relapse mortality (NRM), associated with approximately 20% of deaths for both HLA matched sibling transplants and unrelated donor transplants in 2012-2013 in AML patients (Center for International Blood & Marrow Transplant Research, 2015). The risk of GVHD increases with age; thus, allogeneic HSCT is usually

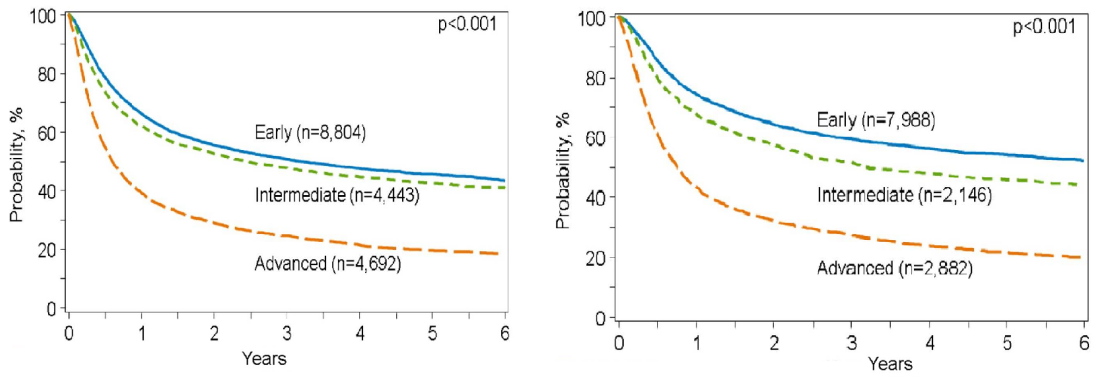

Figure 1.1: Survival of AML patients in early, intermediate and advanced stage of cancer from 2003-2013 after allogeneic HSCT, with unrelated donor (left) and HLA matched sibling donor(right)

restricted to younger patients in good physical condition. GVHD observed within 100 days after HSCT is named acute graft-versus-host disease (aGVHD), which occurs in approximately half of allogeneic HSCT recipients (Ferrara et al., 2009; Weisdorf et al., 2012). AGVHD occurs in the skin, liver, eyes, or gastrointestinal tract once the donor's cells have engrafted in the transplant recipient (Jacobsohn and Vogelsang, 2007). One reason for the high mortality rate associated with aGVHD is that it is difficult to diagnose early and accurately. AGVHD is a clinical diagnosis, mainly based on observed certain symptoms such as fever, skin rash and/or increased dryness, and can be supported with the help of histological confirmation from a biopsy.

At the time of diagnosis, aGVHD is graded by the number and extent of organ involvement. There are two major systems used for grading aGVHD. The first system is the International Bone Marrow Transplant Registry (IBMTR) grading system, which grades severity of aGVHD using the letters A, B, C, and D, with A being least severe and D being most severe. Grading is based upon visual symptoms associated with aGVHD, including rash, diarrhea, and pain. The second system is the Glucksberg grading system, which grades severity of aGVHD using the Roman numerals

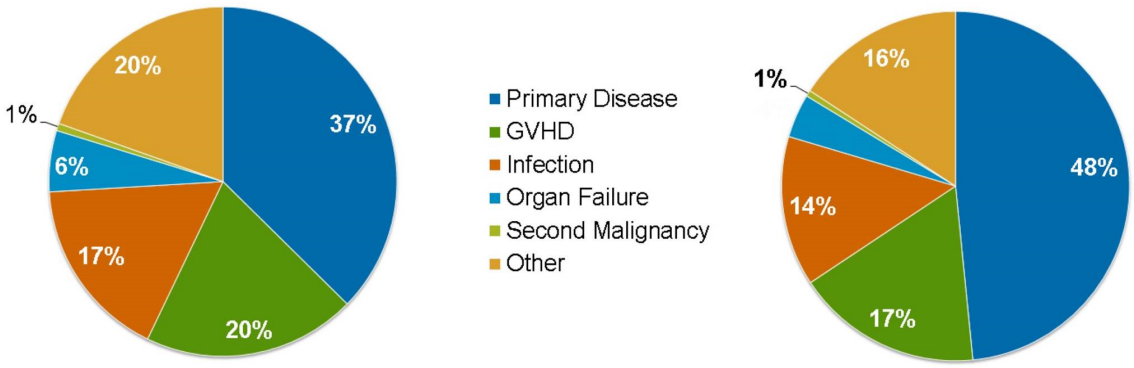

Figure 1.2: Causes of Mortality of AML patients in early, intermediate and advanced stage of cancer from 2003-2013 after allogeneic HSCT, with unrelated donor (left) and HLA-matched sibling donor(right)

I, II, III, and IV, with I indicating least severe and IV indicating most severe. Like IBMTR, Glucksberg includes physical symptoms in its grading. However, Glucksberg also combines a patient's daily living quality of life measured by their Eastern Co-operative Oncology Group (ECOG) score. Treatment is assigned to patients based on their grade of aGVHD severity. Currently steroids remain to be the standard first-line treatment, which can reduce body's immune response and reduce the number of T cells. However, less than 50% of patients will have a complete response to steroids, and steroids have toxic side effects, including osteopenia and immunosuppression (Garnett et al., 2013).

Diagnosis of aGVHD based on clinical symptoms may be inaccurate because some of the symptoms are not specific to aGVHD. Moreover, both clinicians and patients want to avoid invasive biopsy confirmation. Thus, accurate diagnosis and prediction of aGVHD through non-invasive measures are of great importance, because clinicians want to avoid over-treatment and improve patients' quality of life.

1.2 Biomarkers of aGVHD

Much research has been done to explore how biomarkers are related to aGVHD and how they might be used in predicting the onset of aGVHD, NRM, and overall survival (OS) (Paczesny et al., 2009). Candidate biomarkers for the prognosis of aGVHD derive from three general categories: markers of generalized inflammation (e.g., interleukin-8 (IL-8) and tumor necrosis factor-α (TNF-α)), lymphocyte surface molecules (e.g., CD30), and end products secreted from damaged organs (e.g., Elafin and Regenerating islet-derived 3 - α (REG3α)) (Chen and Cutler, 2013). Biomarker levels in plasma often rise several weeks before the clinical disease becomes apparent, making early prediction of the onset of aGVHD plausible (Levine et al., 2006).

Early research includes multiple small studies designed to identify individual blood proteins as biomarkers of aGVHD. Symington et al. (1990) measured the concentration of serum TNF-α in 44 patients who had received HSCT and analyzed the correlation between this concentration and the onset of aGVHD and its severity. The concentration of TNF-α was measured by enzyme-linked immunosorbent assay (ELISA) and dichotimized into two categories: TNF-positive and TNF-negative. Via Fisher's Exact Test, researchers found a weak association between positive levels of TNF-α and aGVHD onset (P = 0.06). Behar et al. (1996) explored whether one poorly defined minor histocompatibility antigen, cluster of differentiation 31 (CD31) adhesion molecule, could explain the high incidence rate of aGVHD among HLA-matched patients. With 46 pairs of recipients of HSCT and their HLA-identical siblings, researchers found that CD31 was a minor alloantigen, and non-identical genotypes of CD31 between donor and recipient was associated with a high risk of aGVHD (P = 0.004). These studies have revealed the value of biomarkers in the prediction of aGVHD, even in small samples of patients. However, in the study by Symington et al. (1990), patients' serum samples were taken between 4 and 52 days post-transplant,

and at the time of serum collection, some of the patients had had developed aGVHD. Biomarkers collected at different times impairs the reliability of this study.

Some researchers suspected that the onset of aGVHD should reflect not only the concentration of biomarkers, but also the change of biomarkers over time. Uguccioni et al. (1993) made serial measurements of serum IL-8 concentration on 8 patients with successful engraftment and 5 patients with aGVHD after HSCT. The IL-8 concentration was measured 20 days before the HSCT, and sequentially after HSCT. Researchers found that the concentration of IL-8 decreased significantly among patients with successful engraftment. However, this IL-8 concentration did not change significantly before and after HSCT among patients developing aGVHD. Another study in 2006 identified that an increase in tumor necrosis factor receptor 1 (TNFR1) on day 7 following allogeneic HSCT compared with its baseline value was strongly correlated with aGVHD onset, NRM and OS (Levine et al., 2006). Both of the two studies were case-control analyses, in which patients were classified as aGVHD patients or aGVHD-free survivors. They used repeatedly collected biomarkers to improve efficiency, however, they lost information by ignoring the true times-to-aGVHD.

Some researchers also argued that differences of any single protein did not have enough specificity and sensitivity to be of clinical use (Paczesny et al., 2009). Thus, researchers have incorporated multiple biomarkers in a multivariate logistic regression model that could hopefully better confirm the diagnosis of aGVHD in patients with onset of clinical symptoms of aGVHD, and provide prognostic information independent of aGVHD severity evaluated based on clinical symptoms (Paczesny et al., 2007, 2009). Harris et al. (2013) found that combining a panel of four biomarkers (IL-2Rα, TNFR1, elafin and REG3α) at day 7 post-HCT and five pre-HCT clinical risk factors produced good prediction of grade II-IV aGVHD following related donor HSCT. This

combination of biomarkers obtained a 77% sensitivity, and clinical factors proved to be significantly more predictive of aGVHD than a model with clinical risk factors only ($P < 0.001$).

Due to the convenience of obtaining biomarkers from plasma, researchers discovered that the prediction of aGVHD onset could be calibrated by a sequential prediction process. Two separate multivariate logistic models were built with IL-2Rα, TNFR1, and elafin at day +7 and day +14, and a new prediction rule was designed using the prediction probabilities of the two models. Patients were labeled as high risk if the predicted probability of aGVHD with day +7 biomarkers was above 0.64. Among the rest low risk group, patients were re-classified as high risk if the predicted probability of aGVHD with day +14 biomarkers was above 0.41 (Paczesny et al., 2011). This approach could be viewed as the first attempt of dynamic prediction of aGVHD.

Moreover, researchers discovered that the severity of symptoms at the onset of aGVHD did not accurately define risk of death, and that most patients were treated similarly with high dose systemic steroids (Levine et al., 2014, 2015). Thus, Harris et al. (2013) built a prognostic score for aGVHD based on TNFR1, REG3α, IL2Rα, elafin and suppressor of tumorigenicity 2(ST2). This new aGVHD grading system based solely on biomarkers reclassified a significant number of patients (n = 21/79, 27%) and produced more accurate risk groups than Glucksberg grades, the most popular grading system of aGVHD based on clinical symptoms, resulting in better NRM prediction as well (Harris et al., 2013; Vander Lugt et al., 2013).

The aforementioned research has proved the efficacy of biomarkers in the prediction and prognosis of aGVHD. Moreover, the sequential prediction model invokes the

need of utilizing repeatedly collected biomarkers to make more accurate prediction of aGVHD. The fundamental theory behind this idea is that though the most recent biomarker values are more related to aGVHD onset than earlier measures, the entire biomarker history offers more information than a single observation. The inexpensive and highly effective ELISA enables biomarkers to be regularly collected. In this project, we will explore various of methods predicting the onset of aGVHD with repeated biomarkers. By doing this, we could achieve aGVHD prediction as early as possible, and refine this prediction whenever a new biomarker observation is available.

1.3 aGVHD Biomarker Dataset

In line with the previous statement, this research is structured for a study conducted by University of Michigan Blood and Marrow Transplant Program. This study includes 381 patients who underwent allogeneic HSCT between the year 2000 and 2010 (Vander Lugt et al., 2013). Their plasma samples were collected weekly throughout the first month, and then monthly thereafter until the first of aGVHD onset or day 100 after HSCT. The concentrations of multiple plasma biomarkers, such as suppression of tumorigenicity 2 (ST2), elafin and IL-2Rα, were measured and recorded. The published data only contain biomarker measurements at day 0, day 14 and day 28 after HSCT, with around 30% missing values at day 14 and day 28. Figure 1.3 shows the concentration of one of the recorded biomarkers, IL2-rα.

In this dataset, patients had at most three repeated biomarker measures, which were insufficient to support a sophisticated model with several parameters. So we simulate data according to this real dataset and evaluate the performance of our methods on the simulated data.

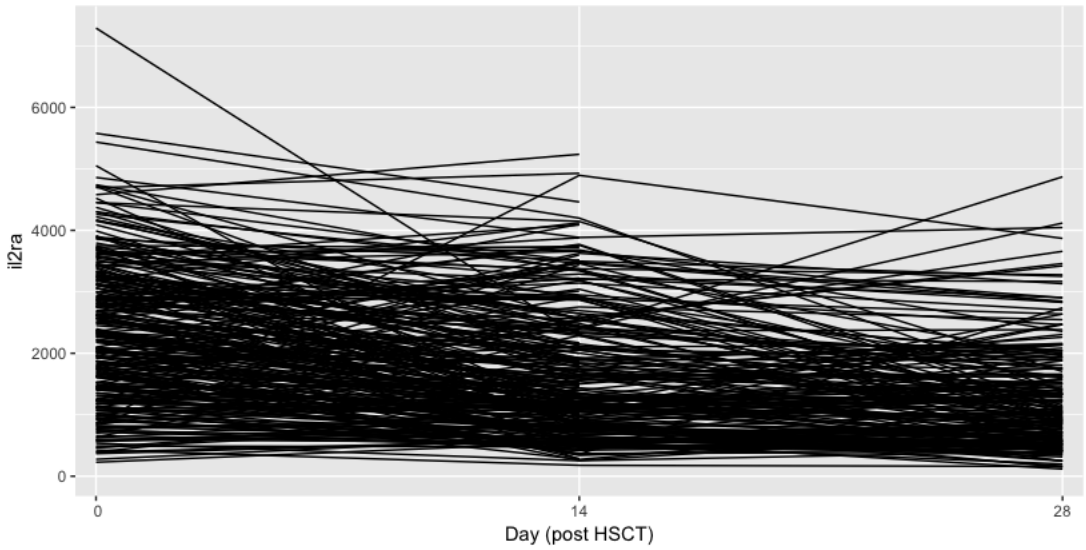

Figure 1.3: IL2-rα levels at day 7, 14, and 28 post HSCT among 381 patients in the University of Michigan Bone Marrow Transplantation program

There are several main features of the simulated data. First, the patients are of various risks of aGVHD, and each aGVHD risk group has a different biomarker trajectory. AGVHD is conventionally defined to occur within 100 days of HSCT, and there are a proportion of patients who will never develop aGVHD within the period of data collection. Moreover, according to the document shared by Center for International Blood & Marrow Transplant Research (2005), there are four identifiable patterns of the frequency trajectory of anti-host T cells, as demonstrated by Figure 1.4. Therefore, we assume patients are from four aGVHD risk groups: high-, medium-, low-risk, and aGVHD-free, with aGVHD risk level-specific mean biomarker profiles. Second, serum biomarker collection tends to be regular and systematic, occurring at specific time intervals after HSCT. Thus, there is little occurrence of missing biomarker values. Third, there is no missingness in time-to-aGVHD for all patients not in the aGVHD-free group. The times-to-aGVHD are observed and recorded for all patients except the aGVHD-free patients.

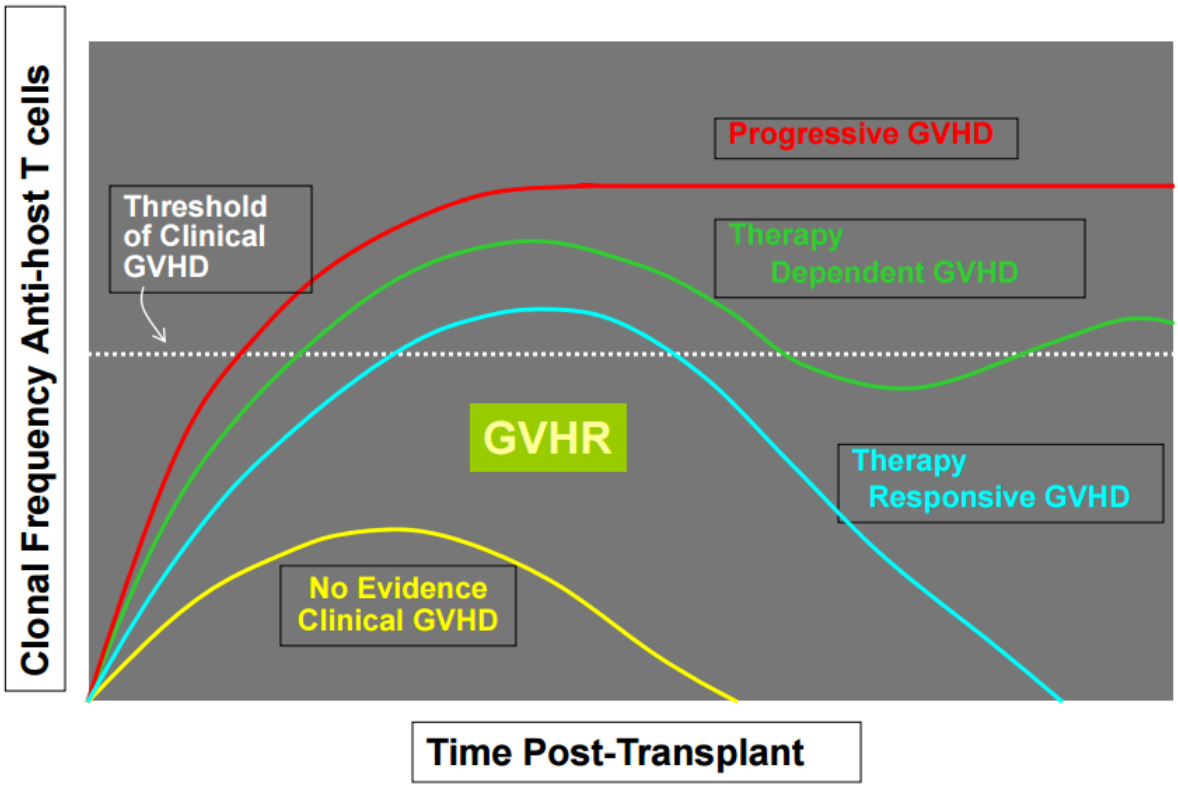

Figure 1.4: the clonal frequency of anti-host T cells over post-transplant period

1.4 Methods for Modeling Longitudinal Processes and Time-to-Event

Modeling a longitudinal process of biomarker observations and a time-to-aGVHD process individually will result in biased estimation of the mean trajectory of biomarkers over time and the hazard ratio quantifying the association between biomarkers and the time-to-aGVHD. This bias is due to the mutual dependence between longitudinal biomarkers and time-to-aGVHD. The time-to-aGVHD depends on the whole history of biomarker levels, while the follow-up of biomarker values is truncated by the time-to-aGVHD (Tsiatis and Davidian, 2004). We now review three existing methods for analyzing times-to-event with repeated biomarkers.

We define $\boldsymbol{Y}_i = (Y_i(t_1), Y_i(t_2), \cdots, Y_i(t_{n_i}))$ as the biomarker history of subject i at time $(t_1, t_2, \cdots, t_{n_i})$, where n_i is the total number of biomarker observations for subject i, for $i = 1, 2, \cdots, n$. Let T_i denote the observed time for subject i, which is the minimum of the time-to-event for subject i, T_i^*, and last follow up time S_i. We also define δ_i as the indicator of whether subject i experiences aGVHD ($\delta_i = 1$) or is censored ($\delta_i = 0$). This notation is used throughout this chapter.

1.4.1 Time-varying Covariates in Times-to-Event

In a survival model with time-varying covariates, the record of subject i is separated into several non-overlapping time intervals $(t_1, t_2), (t_2, t_3), \cdots, (t_{n_i-1}, t_{n_i})$. Within each time interval (t_{j-1}, t_j), the time-varying covariates hold constant values as Y_{j-1}, for $j = 2, 3, \cdots, n_i$. Each interval is coded as one record, thus one subject could have multiple separate records. Standard approaches, i.e., Cox regression, can be applied to this new long-format dataset. At each event time, the risk set is updated, and so are the covariate values.

The key idea behind Cox regression with time-varying covariates is simple: covariates' values are viewed as constant within intervals (Therneau and Lumley, 2011). There are some main drawbacks of using time-varying covariates in the analysis of times-to-event. First, there is no effort to separate the measurement error from true values of time-varying covariates. Second, because biomarkers are only collected intermittently, the exact biomarker value at each event time may not be available. When maximizing the partial likelihood of Cox model with time-varying covariates, the last biomarker observation is carried forward (LOCF). This may lead to biased inference of the association between biomarkers and time-to-event, especially when we have serial multiple biomarkers up to a certain time point, which is shorter than

the total length of follow-up (Fang et al., 2016). Moreover, Cox regression with time-varying covariates is not designed for prediction, because one cannot know how the time-varying covariates change over time. Thus, this approach is not useful in our setting.

1.4.2 Joint Modeling

One popular approach to handle mutually dependent data is joint modeling (Song et al., 2002; Yu et al., 2004; Proust-Lima et al., 2014). Since there is no easy closed-form for the joint distribution of longitudinal biomarkers $\boldsymbol{Y_i}$ and time-to-event (T_i, δ_i), shared terms are brought in to introduce conditional independence between longitudinal process and time-to-event. Two common approaches for expressing shared terms are the latent classes (LC) or shared random effects (SRE), and the joint model is named joint modeling with latent classes (JMLC) and joint modeling with shared random effects (JMSR) correspondingly.

The JMLC and JMSR have fundamentally different assumptions in population heterogeneity of biomarkers and time-to-event distributions. JMSR assumes the population all share the same profile of biomarkers, with the time-to-event depending on individual-level deviations of biomarkers from the population mean. In contrast, JMLC treats subjects as being from different risk groups, and the survival probability only depends on the risk group membership (Blanche et al., 2015; Rizopoulos et al., 2017). In practice, neither assumption fits better in all settings, and the choice between JMLC and JMSR should be made on a case-by-case basis. One major drawback of joint modeling in general is the difficulty in model fitting. Because the LC and SRE are unobservable, they need to be either integrated out of the model, or estimated using either the Expectation-Maximization (EM) algorithm or Markov Chain Monte

Carlo (MCMC).

1.4.3 Landmark Analysis

Another approach, landmark analysis, shows certain benefits over joint modeling with respect to model fitting and interpretation. Zheng and Heagerty (2005) employed landmark analysis for dynamic prediction and defined it as partly conditional modeling. van Houwelingen and Putter (2011) described how to apply landmark analysis to settings with repeated biomarker observations and one final time-to-event endpoint. A standard landmark analysis fits one separate survival model at each predefined landmark time point, so it is easy to execute and straightforward to interpret the parameters.

The main feature of landmark analysis is that at each landmark time, all the future biomarker observations are ignored, and all subjects who have already experienced the event are removed from the risk set. Although landmark analysis is straightforward and easy to execute, it is criticized by its coarse use of biomarker values (van Houwelingen, 2007; van Houwelingen and Putter, 2011). First, the landmark analysis abandons the whole biomarker history before the landmark time and only uses the biomarker observations at the landmark time. When a biomarker value is missing, the last observed biomarker value is carried forward to the landmark time. Second, a landmark analysis ignores the measurement error of biomarker values and fits the time-to-event model without adjusting for this noise. If the longitudinal biomarkers are measured sparsely and irregularly, and also with measurement errors, landmark analysis might not be an ideal approach.

1.5 Structure of Dissertation

In the second chapter, we analyze longitudinal biomarkers and time-to-aGVHD with both JMLC and landmark analysis. Both methods are modified specifically to our setting. We introduce how to do model fitting and prediction with these two methods, and compare their prediction performance. Though these two approaches are well-accepted, both JMLC and landmark analysis have complex model specification and the model fitting is time consuming. Thus, in the third chapter, we build a pattern mixture model to predict the onset of aGVHD given longitudinal biomarker values. This pattern mixture model is easy to execute and interpret. In the fourth chapter, we generalized this pattern mixture model to incorporate censored cases. The JMLC model used in the second chapter requires a pre-specified number of latent classes, thus, in the fifth chapter, we propose a hypothesis testing based model selection process to select the number of latent classes in JMLC. The future research areas that can be explored further based on our current work are discussed after.

CHAPTER II

Dynamic Prediction of Time-to-acute Graft-versus-Host-Disease with Joint Modeling and Landmark Analysis

2.1 Introduction

In this chapter, we use two methods to assess future risk of aGVHD based on repeated biomarker observations. Our first approach uses a revised JMLC, which includes one latent class for those who will never develop aGVHD (aGVHD-free), and several other classes defined by the risk of aGVHD. Given that one patient's time-to-aGVHD is beyond day 100, his/her aGVHD-free class membership is labeled at the beginning in order to increase model identifiability. Our second approach applies a landmark analysis that is modified to allow for patients from various aGVHD risk groups.

These two methods both require a pre-specified number of risk classes among the patients. This number can be a subjective choice made based on data visualization or the clinicians' prior medical knowledge of aGVHD. Moreover, previous studies have also discussed choosing the number of latent classes based on the model selection results. For example, in a study of using prostate-specific antigen (PSA) to predict the

risk of prostate cancer, a separate model was fitted for each of a varying number of latent classes; the model with the least value of Bayesian information criterion (BIC) was chosen as the final model (Lin et al., 2002).

For the purposes of this project, we will fix the number of latent classes according to the medical characteristics of aGVHD patients. According to Center for International Blood & Marrow Transplant Research (2005), clinicians have labeled aGVHD as no-evidence clinical aGVHD, therapy-responsive aGVHD, therapy-dependent aGVHD, and progressive aGVHD. Moreover, based on clinical symptoms, aGVHD can be also classified into four severity phases, according to the two major systems used for grading aGVHD, IBMTR and Glucksberg grading system. Thus, in this study, we will assume the patients come from four latent aGVHD groups, which equals the true inherent number of latent aGVHD groups in the simulated data. More discussions on choosing the number of latent classes based on the model selection results can be found in Chapter V.

One distinctive feature of aGVHD data is that a subset of patients are "aGVHD-free." Based on the clinical properties of allogeneic bone marrow transplantation (BMT) and disease definition, patients who have not developed aGVHD within 100 days after BMT are assumed to never develop aGVHD. Thus, we classify all patients who have not experienced aGVHD within the first 100 days after BMT in the "aGVHD-free" latent class.

The rest of this chapter is organized as follows. First, we introduce how to do model fitting and prediction in JMLC. Next, we talk about the modeling fitting and prediction of aGVHD in landmark analysis, followed by the metrics for evaluating aGVHD predictions. Simulations under different scenarios are executed to check the

prediction performance of these two methods under various scenarios. We conclude with a brief discussion.

2.2 Joint Modeling with Latent Class

2.2.1 Model Setting

JMLC is proposed to model outcomes of various types, especially when there is no closed-form for the joint distribution for these outcomes. In aGVHD data, the outcome is a combination of repeated biomarker observations and times-to-aGVHD, and we assume the distributions of biomarkers and times-to-aGVHD are different across latent risk groups of aGVHD. To establish JMLC, we first define some notations used throughout this section.

Let $\boldsymbol{z_i} = (z_{i1}, z_{i2}, z_{i3}, z_{i4})'$ represent the unobserved indicator vector of subject i's latent class membership, where $z_{ih} = 1$ if subject i belongs to latent class h, and 0 otherwise, for $h = 1, 2, 3,$ and 4. Here we fix z_{i1} as the indicator of the aGVHD-free latent class. Let $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \pi_{i3}, \pi_{i4})'$ be the corresponding probabilities, and $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)'$ be the marginal probabilities of one subject belonging to each aGVHD class, with the restriction that $\sum_{h=1}^{4} \pi_h = 1$.

For the longitudinal biomarker process, we define $\boldsymbol{Y}_i = (Y_i(t_1), Y_i(t_2), \cdots, Y_i(t_{n_i}))$ as the biomarker history of subject i at times $(t_1, t_2, \cdots, t_{n_i})$, where n_i is the total number of biomarker observations for subject i, for $i = 1, 2, \cdots, n$. We assume the biomarker data are collected according to a medical protocol based on the features of aGVHD and time feasibility. Therefore, the timing of biomarker screening is unrelated to patients health conditions, or more specifically, patients' biomarker levels.

We specify that patients from the same latent class of aGVHD share the same mean biomarker trajectory, with individual-specific random effects $\boldsymbol{b_i}$ reflecting the deviation of an individual's biomarker pattern from the mean of their latent class. The measurement error $\boldsymbol{e_i} = (e_i(t_1), e_i(t_2), \cdots, e_i(t_{n_i}))$ of biomarkers introduces the random noise in biomarker measurement. We assume the observed biomarkers, $\boldsymbol{Y_i}$, random effects, $\boldsymbol{b_i}$, and measurement error, $\boldsymbol{e_i}$, $\boldsymbol{B_i}' = (\boldsymbol{Y_i}, \boldsymbol{b_i}, \boldsymbol{e_i})$ given $z_{ih} = 1$ have a multivariate normal distribution, i.e.

$$\boldsymbol{B_i}|z_{ih} = 1 \sim \mathcal{MVN}\left(\begin{pmatrix} X_i\boldsymbol{\beta^{(h)}} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} Z_iDZ_i^T + \sigma^2\boldsymbol{I_{n_i}} & Z_iD & \sigma^2\boldsymbol{I_{n_i}} \\ DZ_i^T & D & \mathbf{0} \\ \sigma^2\boldsymbol{I_{n_i}} & \mathbf{0} & \sigma^2\boldsymbol{I_{n_i}} \end{pmatrix} \right) \tag{2.1}$$

with a density function $f_h(\boldsymbol{B_i})$, where X_i is the design matrix including functions of time, $\boldsymbol{\beta^{(h)}}$ is the corresponding parameters of the mean biomarker trajectory in the latent class h, Z_i is design matrix of random effects that can be any subset of X_i, D is the covariance matrix of the random effects that is constant among all latent classes, and σ^2 is the common variance of each element of $\boldsymbol{e_i}$. Let $\boldsymbol{\omega} = (\boldsymbol{\beta^{(1)}}, \boldsymbol{\beta^{(2)}}, \boldsymbol{\beta^{(3)}}, \boldsymbol{\beta^{(4)}}, D, \sigma^2)$ denote all the parameters involved in the longitudinal process.

For the time-to-aGVHD process, let T_i denote the observed time for subject i, which is the minimum of time-to-aGVHD for subject i, T_i^*, and last follow-up time S_i. We also define δ_i as the indicator of whether subject i experiences aGVHD ($\delta_i = 1$) or is censored ($\delta_i = 0$). For the "cured" latent class, the hazard of aGVHD is 0 and the aGVHD probability is 0. For simplicity, we assume a Weibull distribution for the time-to-aGVHD of the three other latent classes, i.e., $Pr(T_i > t|i \in h) = exp(-(t/\lambda^{(h)})^{\kappa^{(h)}})$, with hazard function $g^{(h)}(t) = k^{(h)}/\lambda^{(h)}(t/\lambda^{(h)})^{(\kappa^{(h)}-1)}exp(-(t/\lambda^{(h)})^{\kappa^{(h)}})$, for $h = 2, 3$ and 4. Here we define $\boldsymbol{\zeta} = (\boldsymbol{\lambda} = (\lambda^{(2)}, \lambda^{(3)}, \lambda^{(4)}), \boldsymbol{\kappa} = (\kappa^{(2)}, \kappa^{(3)}, \kappa^{(4)}))$ to

be the parameters involved in the time-to-aGVHD process.

Let $\boldsymbol{\xi} = (\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{\zeta})$ represent the complete parameter set; the complete data are $(\boldsymbol{Y_i}, \boldsymbol{b_i}, T_i, \delta_i, \boldsymbol{z_i})$. The complete data likelihood function is:

$$\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{Y}, \boldsymbol{b}, T, \delta, \boldsymbol{z}) = \prod_{i=1}^{n} \left[[\pi_1 f_1(\boldsymbol{B_i}|\boldsymbol{w})]^{z_{i1}} \prod_{h=2}^{4} [\pi_h f_h(\boldsymbol{B_i}|\boldsymbol{w})(g^{(h)}(T_i))^{\delta_i} exp(-\Big(\frac{T_i}{\lambda^{(h)}}\Big)^{\kappa^{(h)}})]^{z_{ih}} \right]$$

with corresponding log-likelihood:

$$l(\boldsymbol{\xi}|\boldsymbol{B}, T, \delta, \boldsymbol{z}) = l_1(\boldsymbol{\pi}|\boldsymbol{B}, T, \delta, \boldsymbol{z}) + l_2(\boldsymbol{\omega}|\boldsymbol{B}, T, \delta, \boldsymbol{z}) + l_3(\boldsymbol{\zeta}|\boldsymbol{B}, T, \delta, \boldsymbol{z}) \tag{2.2}$$

where $l_1(\boldsymbol{\pi}|\boldsymbol{B}, T, \delta, \boldsymbol{z}) = \sum_{i=1}^{n} \sum_{h=1}^{4} z_{ih} log(\pi_h)$, $l_2(\boldsymbol{\omega}|\boldsymbol{B}, T, \delta, \boldsymbol{z}) = \sum_{i=1}^{n} \sum_{h=1}^{4} z_{ih} log f_h(\boldsymbol{B_i}|\boldsymbol{\omega})$, and $l_3(\boldsymbol{\zeta}|\boldsymbol{B}, T, \delta, \boldsymbol{z}) = \sum_{i=1}^{n} \sum_{h=2}^{4} z_{ih}[\delta_i log(g^{(h)}(T_i)) - (T_i/\lambda^{(h)})^{\kappa^{(h)}}]$, which are the three components of the log-likelihood corresponding to $\boldsymbol{\pi}, \boldsymbol{\omega}$ and $\boldsymbol{\zeta}$. Because we cannot observe the individual-level latent class indicator $\boldsymbol{z_i}$ in practice, we use the Expectation-Maximization (EM) algorithm to find the expectation of unobserved $\boldsymbol{z_i}$ and maximum likelihood estimate (MLE) of parameters iteratively. Section 2.2.2 describes the details of using the EM algorithm to maximize the aforementioned log-likelihood in Equation 2.2.

2.2.2 Parameter Estimation with the EM Algorithm

In the E-step, at iteration $q+1$, we estimate the expectation of unobserved complete-data sufficient statistics, $(\boldsymbol{z_i}, \boldsymbol{b_i b_i'}, \boldsymbol{e_i e_i'})$, conditionally on the parameter estimates $\boldsymbol{\xi}^{(q)}$ from the previous iteration q. Here $E(\boldsymbol{z_i}|\boldsymbol{\xi}^{(q)}, \boldsymbol{Y_i}, T_i, \delta_i) = \pi_{ih}(\boldsymbol{\xi}^{(q)})$ is the probability subject i at iteration $q+1$ belongs to latent class h. For simplicity, throughout the rest of subsections 2.2.2 and 2.2.3 we will use $\boldsymbol{\xi}$ instead of $\widehat{\boldsymbol{\xi}^{(q)}}$

to represent the parameter estimation at iteration q. For the cases with observed times-to-aGVHD, $\pi_{ih}(\boldsymbol{\xi})$ is computed as:

$$\pi_{ih}(\boldsymbol{\xi}) = \frac{\pi_h f_h(\boldsymbol{Y_i}|\boldsymbol{\omega})(g^h(T_i))^{\delta_i} exp(-(t/\lambda^{(h)})^{\kappa^{(h)}}}{\sum_{l=2}^{4} \pi_l f_l(\boldsymbol{Y_i}|\boldsymbol{\omega})(g^l(T_i))^{\delta_i} exp(-(t/\lambda^{(l)})^{\kappa^{(l)}})} \tag{2.3}$$

for $h = 2$, 3, and 4. The value of $\pi_{i1}(\boldsymbol{\xi})$ for aGVHD-free patients is fixed at 1.

Next, we compute the expectation of $\boldsymbol{b_i b_i'}$ and $\boldsymbol{e_i e_i'}$. Define $\boldsymbol{H} = Z_i D Z_i' + \sigma^2 \boldsymbol{I_{n_i}}$, and given the joint distribution of $\boldsymbol{B_i}$ we obtain:

$$\begin{aligned}
E(\boldsymbol{b_i b_i'}|\boldsymbol{Y_i}, \boldsymbol{\xi}) &= E_{\boldsymbol{b_i}|\boldsymbol{Y_i}}(\boldsymbol{b_i}|\boldsymbol{Y_i}) E_{\boldsymbol{b_i}|\boldsymbol{Y_i}}(\boldsymbol{b_i}|\boldsymbol{Y_i})' + cov_{\boldsymbol{b_i}|\boldsymbol{Y_i}}(\boldsymbol{b_i}|\boldsymbol{Y_i}) \\
E(\boldsymbol{e_i e_i'}|\boldsymbol{Y_i}, \boldsymbol{\xi}) &= E_{\boldsymbol{e_i}|\boldsymbol{Y_i}}(\boldsymbol{e_i}|\boldsymbol{Y_i}) E_{\boldsymbol{e_i}|\boldsymbol{Y_i}}(\boldsymbol{e_i}|\boldsymbol{Y_i})' + tr\{cov_{\boldsymbol{e_i}|\boldsymbol{Y_i}}(\boldsymbol{e_i}|\boldsymbol{Y_i})\} \\
\text{where} \quad E_{\boldsymbol{b_i}|\boldsymbol{Y_i}}(\boldsymbol{b_i}|\boldsymbol{Y_i}) &= D Z_i \boldsymbol{H}^{-1}(\boldsymbol{Y_i} - X_i \boldsymbol{\beta}^{(h)}) \\
cov_{\boldsymbol{b_i}|\boldsymbol{Y_i}}(\boldsymbol{b_i}|\boldsymbol{Y_i}) &= D - D Z_i \boldsymbol{H}^{-1} Z_i D \\
E_{\boldsymbol{e_i}|\boldsymbol{Y_i}}(\boldsymbol{e_i}|\boldsymbol{Y_i}) &= \sigma^2 \boldsymbol{H}^{-1}(\boldsymbol{Y_i} - X_i \boldsymbol{\beta}^{(h)}) \\
cov_{\boldsymbol{e_i}|\boldsymbol{Y_i}}(\boldsymbol{b_i}|\boldsymbol{Y_i}) &= \sigma^2(\boldsymbol{I_{n_i}} - \sigma^2 \boldsymbol{H}^{-1})
\end{aligned}$$

With the complete data sufficient statistics $(\boldsymbol{z_i}, \boldsymbol{b_i b_i'}, \boldsymbol{e_i e_i'})$, we can compute the three components, $l_1(\boldsymbol{\pi}|\boldsymbol{B}, T, \delta, \boldsymbol{z})$, $l_2(\boldsymbol{\omega}|\boldsymbol{B}, T, \delta, \boldsymbol{z})$, and $l_3(\boldsymbol{\zeta}|\boldsymbol{B}, T, \delta, \boldsymbol{z})$, in the expectation of log-likelihood in Equation 2.2. In the M-step, we can compute the MLE for parameters $\boldsymbol{\xi} = (\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{\zeta})$ by maximizing the corresponding expectation of the log-likelihood. The MLE for the parameters in the longitudinal biomarker process,

$\boldsymbol{\omega}$, are:

$$\begin{aligned}
\widehat{D} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{4}\pi_{ih}(\boldsymbol{\xi})\Big[E_{\boldsymbol{b_i}|\boldsymbol{Y_i}}(\boldsymbol{b_i}|\boldsymbol{Y_i})E_{\boldsymbol{b_i}|\boldsymbol{Y_i}}(\boldsymbol{b_i}|\boldsymbol{Y_i})^T + cov_{\boldsymbol{b_i}|\boldsymbol{Y_i}}(\boldsymbol{b_i}|\boldsymbol{Y_i})\Big] \\
\widehat{\sigma^2} &= \frac{1}{\sum_{i=1}^{n} n_i}\sum_{i=1}^{n}\sum_{h=1}^{4}\pi_{ih}(\boldsymbol{\xi})\Big[E_{\boldsymbol{e_i}|\boldsymbol{Y_i}}(\boldsymbol{e_i}|\boldsymbol{Y_i})E_{\boldsymbol{e_i}|\boldsymbol{Y_i}}(\boldsymbol{e_i}|\boldsymbol{Y_i})^T + tr\{cov_{\boldsymbol{e_i}|\boldsymbol{Y_i}}(\boldsymbol{e_i}|\boldsymbol{Y_i})\}\Big] \\
\widehat{\boldsymbol{\beta^{(h)}}} &= \Big[\sum_{i=1}^{n}\pi_{ih}(\boldsymbol{\xi})X_i^T\boldsymbol{H_i}^{-1}X_i\Big]^{-1}\Big[\sum_{i=1}^{n}\pi_{ih}(\boldsymbol{\xi})X_i^T\boldsymbol{H_i}^{-1}Y_i\Big]
\end{aligned}$$

The MLE of $\boldsymbol{\zeta}$, the parameters in the time-to-aGVHD process, can be estimated by maximizing $E(l_3(\boldsymbol{\zeta}|\boldsymbol{B}, T, \delta, \boldsymbol{z}))$. Since there is no closed form for the MLE, we will use Newton's method to find the MLE iteratively. Moreover, several existing packages in R offer model fitting procedures for survival data with weights (Therneau, 2015). The MLE of $\boldsymbol{\pi}$ is $\pi_h = \sum_{i=1}^{n}\pi_{ih}(\boldsymbol{\xi})/n$.

2.2.3 Prediction of Time-to-aGVHD

A patient receiving HSCT is scheduled to have serum drawn each week from which biomarkers are measured. This procedure continues until this patient develops aGVHD or reaches 100 days without aGVHD. The prediction of aGVHD onset is made after we obtain two biomarker observations, and this prediction is updated every week when a new biomarker observation is available.

At week k, we obtain biomarker values $\boldsymbol{Y_m}(k) = (Y_{m1}, Y_{m2}, \cdots, Y_{mk})$ of a new patient m, who inherently is aGVHD-free before week k. The probability that patient

m will not develop aGVHD for the next two weeks is:

$$Pr(T_m > k+2|T_m > k) = \sum_{h=2}^{4} Pr(z_{mh}=1|\boldsymbol{Y_m}(k))Pr(T_m > k+2|T_m > k, m \in h) + Pr(z_{m1}=1|\boldsymbol{Y_m}(k))$$

$$\text{where} \quad Pr(z_{mh}=1|\boldsymbol{Y_m}(k)) = \frac{\pi_h f_h(\boldsymbol{Y_m}|\boldsymbol{\omega})P(T_m > k|m \in h, \boldsymbol{\zeta})}{\sum_{l=2}^{4} \pi_l f_l(\boldsymbol{Y_m}|\boldsymbol{\omega})P(T_m > k|m \in l, \boldsymbol{\zeta})},$$

and $Pr(T_m > k|z_{m1} = 1) = 1$ for any k. So the probability of not developing aGVHD in the next two weeks is the probability of patient m falls into aGVHD-free class, plus a weighted sum of probabilities of not developing aGVHD for the next two weeks, with weights equal to the probability that patient m belongs to each risk class.

2.3 Landmark Analysis

2.3.1 Model Setting

In a landmark analysis, a series of fixed times $s = (s_1, s_2, \cdots, s_R)$ after HSCT are selected as the landmark times. Unlike JMLC, which fits one overall model with all the available data, landmark analysis updates the risk set and fits a separate model at each landmark time. We ignore the patients who have developed aGVHD or been censored before the landmark time. For patients developing aGVHD after the landmark time, we ignore biomarkers measured after the landmark time. A natural choice of landmark time here would be the weekly biomarker screening day, because that is when the biomarker information is updated, and the timing of biomarker screening is independent of patients' current or past biomarker levels.

As in Section 2.2, we define $\boldsymbol{Y}_i = (Y_i(t_1), Y_i(t_2), \cdots, Y_i(t_{n_i}))$ to be the biomarker history of subject i at times $(t_1, t_2, \cdots, t_{n_i})$, where n_i is the total number of biomarker

observations for subject i, for $i = 1, 2, \cdots, n$. Let T_i denote the observed time for subject i, which is the earlier of the time-to-aGVHD for subject i, T_i^*, and the last follow-up time, S_i. We also define δ_i as the indicator of whether subject i experiences aGVHD ($\delta_i = 1$) or is censored ($\delta_i = 0$). Since the biomarker observations of all patients are balanced and equally spaced at each landmark time, we can cluster the samples without specifying the functional form of the biomarker trajectory, possibly avoiding biased results caused by model misspecification.

Instead, we assume these biomarker observations are from a mixture of multivariate normal distributions, with the mixture defined by the membership in each aGVHD risk class. Define v_r as the number of biomarker samples by landmark time s_r. The distribution of biomarkers $\boldsymbol{Y_i} = (Y_{i1}, Y_{i2}, \cdots, Y_{i,v_r})$ for subject i who is still at risk for aGVHD is:

$$\boldsymbol{Y_i} \sim \sum_{h=1}^{4} \pi_h^{(s_r)} \mathcal{N}(\boldsymbol{\mu_h}^{(s_r)}, \Sigma_h^{(s_r)}) \tag{2.4}$$

where $\boldsymbol{\mu_h}^{(s_r)} = (\mu_{h1}, \mu_{h2}, \cdots, \mu_{h,v_r})$ is the mean profile of biomarkers in class h at landmark time s_r, $\Sigma_h^{(s_r)}$ is the corresponding variance-covariance matrix, and $\boldsymbol{\pi}^{(s_r)} = (\pi_1^{(s_r)}, \pi_2^{(s_r)}, \pi_3^{(s_r)}, \pi_4^{(s_r)})$ is the marginal probabilities of a patient belonging to each class at landmark time s_r. In a landmark analysis, we allow the marginal probabilities of class membership $\boldsymbol{\pi}$ change along landmark time s.

Additional assumptions on $\boldsymbol{\mu_h}^{(s_r)}$ and $\Sigma_h^{(s_r)}$ can be made to reduce the number of parameters to estimate. For example, in this study we assume $\Sigma_h^{(s_r)} = \Sigma^{(s_r)}$ across all classes. Moreover, one can assume an AR(1) structure of covariance between biomarkers from the same subject. For simplicity, through the rest of this subsection we will omit the superscript of landmark time s_r.

Because the patients at risk for aGVHD are updated at each landmark time, the mean and variance-covariance of biomarkers are updated at each landmark time. The likelihood for the parameters $\boldsymbol{\xi} = (\boldsymbol{\mu}, \Sigma, \boldsymbol{\pi})$ at landmark time s_r is:

$$\mathcal{L}^{(s_r)}(\boldsymbol{\xi}|\boldsymbol{Y}, \boldsymbol{z}) = \prod_{i=1}^{N_{s_r}} \prod_{h=1}^{4} \{\pi_h |\Sigma|^{-\frac{1}{2}} exp(-\frac{1}{2}(\boldsymbol{Y_i} - \boldsymbol{\mu_h})' \Sigma^{-1} (\boldsymbol{Y_i} - \boldsymbol{\mu_h}))\}^{z_{ih}}$$

where $\boldsymbol{z_i} = (z_{i1}, z_{i3}, z_{i4}, z_{i4})^T$ represents the indicator vector of subject i's latent class membership, and N_{s_k} is number of patients at risk at landmark time s_k. The corresponding log-likelihood of $\boldsymbol{\xi}$ is:

$$l^{(s_r)}(\boldsymbol{\xi}|\boldsymbol{Y}, \boldsymbol{z}) = \sum_{i=1}^{N_{s_r}} \sum_{h=1}^{4} \Big[z_{ih} log(\pi_h) - \frac{z_{ih} log(|\Sigma|)}{2} - \frac{z_{ih}}{2} (\boldsymbol{Y_i} - \boldsymbol{\mu_h})' \Sigma^{(s_k)-1} (\boldsymbol{Y_i} - \boldsymbol{\mu_h}) \Big] \quad (2.5)$$

The group indicator of each individual, $\boldsymbol{z_i}$, is unobservable, so we will employ the EM algorithm iteratively to compute the expectation of $\boldsymbol{z_i}$ and the MLE of $\boldsymbol{\xi}$ at each landmark time s_r. We will describe the details of using the EM algorithm to compute the MLE of $\boldsymbol{\xi}$ in subsection 2.3.2.

2.3.2 Parameter Estimation with the EM Algorithm

At iteration q and in the E-step, the expectation of $\boldsymbol{z_{ih}}$ is calculated as:

$$\pi_{ih}(\boldsymbol{\xi}^{(q)}) \quad = \quad \frac{\pi_h^{(q)} exp(-\frac{1}{2}(\boldsymbol{Y_i} - \boldsymbol{\mu_h}^{(q)})' \Sigma^{-1(q)} (\boldsymbol{Y_i} - \boldsymbol{\mu_h}^{(q)}))}{\sum_{l=1}^{4} \pi_l^{(q)} exp(-\frac{1}{2}(\boldsymbol{Y_i} - \boldsymbol{\mu_l}^{(q)})' \Sigma^{-1(q)} (\boldsymbol{Y_i} - \boldsymbol{\mu_l}^{(q)}))} \quad (2.6)$$

Then in the M-step, by maximizing the expectation of the log-likelihood as in Equation 2.5, we achieve the MLE of $\boldsymbol{\xi} = (\boldsymbol{\mu_h}, \Sigma, \boldsymbol{\pi})$ as:

$$\begin{aligned}
\widehat{\pi_h} &= \frac{1}{N_{s_r}} \sum_{i=1}^{N_{s_r}} \pi_{ih}(\widehat{\boldsymbol{\xi}^{(q)}}) \\
\widehat{\boldsymbol{\mu_h}} &= \frac{\sum_{i=1}^{N_{s_r}} \pi_{ih}(\widehat{\boldsymbol{\xi}^{(q)}}) \boldsymbol{Y_i}}{\sum_{i=1}^{N_{s_r}} \pi_{ih}(\widehat{\boldsymbol{\xi}^{(q)}})} \\
\widehat{\Sigma} &= \frac{\sum_{i=1}^{N_{s_r}} \sum_{h=1}^{4} \pi_{ih}(\widehat{\boldsymbol{\xi}^{(q)}})(\boldsymbol{Y_i} - \widehat{\boldsymbol{\mu_h^q}})(\boldsymbol{Y_i} - \widehat{\boldsymbol{\mu_h^q}})'}{N_{s_r}}
\end{aligned} \tag{2.7}$$

When convergence criteria are met, we achieve a set of parameter estimates of $\widehat{\boldsymbol{\xi}} = (\widehat{\boldsymbol{\mu_h}}, \widehat{\Sigma}, \widehat{\boldsymbol{\pi}})$, as well as the individual probability of belonging to each risk class $\widehat{\boldsymbol{\pi_i}} = (\widehat{\pi_{i1}}, \widehat{\pi_{i2}}, \widehat{\pi_{i3}}, \widehat{\pi_{i4}})$.

Since landmark analysis ignores all the biomarkers observed after the landmark time, the accuracy of long-term prediction is reduced. In practice, we will report two-week prediction of aGVHD onset, so we will explore the two-week onset rate of aGVHD in the model fitting stage.

We recode the time-to-aGVHD into an indicator W_i, representing whether or not this individual experiences aGVHD within the next two weeks, with $W_i = 1$ representing aGVHD onset and $W_i = 0$ representing no aGVHD. We assume the probability of experiencing aGVHD follows a binomial distribution. The logit of this probability for patients in risk group h is modeled as:

$$log\left(\frac{Pr(W_i = 1 | z_{ih} = 1)}{1 - Pr(W_i = 1 | z_{ih} = 1)}\right) = \alpha^h \tag{2.8}$$

where $\boldsymbol{\alpha} = (\alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)}, \alpha^{(4)})$ denotes the marginal log odds ratio of developing

aGVHD in two weeks in each aGVHD latent class. The log-likelihood of $\boldsymbol{\alpha}$ is:

$$l(\boldsymbol{\alpha}|\boldsymbol{W}, \boldsymbol{\pi_i}) = \sum_{h=1}^{4} \left(\sum_{i=1}^{N_{s_r}} \pi_{ih}[\alpha^{(h)} W_i - log(1 + exp(\alpha^{(h)}))] \right) \tag{2.9}$$

Then we can obtain the MLE of $\boldsymbol{\alpha}$ by maximizing Equation 2.9. However, there is no closed form for the MLE of $\boldsymbol{\alpha}$ from Equation 2.9, so we will use Newton's method and get the MLE iteratively.

2.3.3 Prediction in Landmark Analysis

At week k, we want to make a prediction of aGVHD onset for a new patient m, who has biomarker observation history $\boldsymbol{Y_m}(k) = (Y_{m1}, Y_{m2}, \cdots, Y_{mk})$, and the patient m is inherently aGVHD-free before week k. The probability of being aGVHD-free for the next two weeks is:

$$Pr(T_m > k + 2 | T_m > k) = \sum_{h=1}^{4} \pi_{mh} \frac{exp(\widehat{\alpha^h})}{1 + \widehat{exp(\alpha^h)}} \tag{2.10}$$

$$\text{where } \pi_{mh} = \frac{\widehat{\pi_h} exp(-\frac{1}{2}(\boldsymbol{Y_m} - \widehat{\boldsymbol{\mu_h}})' \widehat{\Sigma}^{-1} (\boldsymbol{Y_m} - \widehat{\boldsymbol{\mu_h}}))}{\sum_{l=1}^{4} \widehat{\pi}_l exp(-\frac{1}{2}(\boldsymbol{Y_m} - \widehat{\boldsymbol{\mu_l}})' \widehat{\Sigma}^{-1} (\boldsymbol{Y_m} - \widehat{\boldsymbol{\mu_l}}))}$$

The MLE of parameters $(\boldsymbol{\xi}, \boldsymbol{\alpha})$ achieved at landmark time s_k are used in Equation 2.10.

2.4 Evaluation of Prediction

The evaluation of prediction falls into two fields, discrimination and calibration. Discrimination indexes measure how well the model distinguishes between patients who experience aGVHD from patients who do not experience aGVHD (Schoop et al., 2011; Blanche et al., 2015; Yang et al., 2016; Rizopoulos et al., 2017). Calibration indexes evaluate how accurate the model predicts the probability of aGVHD (Schoop et al., 2011; Rizopoulos et al., 2017).

In our simulation study, we will evaluate the prognosis of aGVHD onset at each prediction time with both Brier Score (BS) and area under the curve (AUC). BS assesses the absolute accuracy of predictions, and it has been widely used in evaluation of prediction performance in survival analysis (Brier, 1950; Schemper and Henderson, 2000; Henderson et al., 2002; Rizopoulos et al., 2017). In our study, we define the two-week BS at week k as:

$$BS(k) = \sum_{i=1}^{N_k} (I_i(k+2) - Pr_i(k+2))^2 \tag{2.11}$$

where N_k is the number of patients at risk at week k , and $I_i(k+2)$ and $Pr_i(k+2)$ are the respective indicator of aGVHD status and predicted probability of aGVHD at week $k+2$ for patient i. Lower values of BS indicate better prediction, with a perfect prediction indicated by $BS = 0$.

Alternatively, we use a receiver operating characteristic (ROC) curve to evaluate the discrimination ability of our models. For each cut-off point of predicted probabilities, the resulting sensitivity and specificity are indicated as a point on the curve. The AUC is introduced to summarize the overall discrimination performance of a prediction model, with a value of 0.5 indicating no predictive ability and a value of 1 indicating a perfect discrimination.

2.5 Simulation and Result

We speculate that several factors might influence the relative performance of JMLC and landmark analysis. The first factor of practical interest is the overlap of time-to-aGVHD distributions between latent classes. The second factor is the size of biomarker measurement error. The third factor is the assumed functional

form of the biomarker patterns over time. In order to examine the effects of these three factors, we generate more or less overlapping distributions of time-to-aGVHD between latent classes, add large or small measurement errors to biomarker observations, and assume a linear trajectory of biomarkers over time or other irregular forms.

In our simulations, biomarker screening is scheduled right after HSCT (baseline) and weekly thereafter until the onset of aGVHD. An uninformative baseline biomarker level is assumed, so at least two biomarker observations are needed to make a prediction for aGVHD onset. In each simulation, a sample of 200 patients is generated as the training dataset, and another sample of 200 patients from the same population is generated as the testing dataset. This population consists of subjects from four latent classes: aGVHD-free, low-risk, medium-risk, and high-risk. Patients within the same latent class share the same distribution of time-to-aGVHD and mean biomarker profile. With each testing dataset, predictions are made at week $1, 2, \cdots, 8$ of aGVHD onset within the next two weeks. We start from the end of week 1 because this is when two biomarker observations are available for each patient, and we end at week 8 because a majority of aGVHD incidence occurs within 10 weeks of HSCT. The results are based on 5,000 simulations.

2.5.1 Effect of Overlap in Time-to-aGVHD Distributions between Latent Classes

First, we examine the impact of overlap in times-to-aGVHD between latent classes. This overlap reflects the variance of the times-to-aGVHD in each latent class. If there is less overlap of times-to-aGVHD between latent classes, patients with similar biomarker patterns are more likely to have similar times-to-aGVHD. Thus, the distribution of times-to-aGVHD in one latent class, in which patients share same

biomarker pattern, is concentrated. Otherwise, if there is extensive overlap of the time-to-aGVHD distributions between latent classes, it is hard to tell one patient's latent class membership given only his/her biomarker profile.

We simulate data from two degrees of overlap in time-to-aGVHD distribution, as demonstrated by Figure 2.1. When there is less overlap in time-to-aGVHD distri-

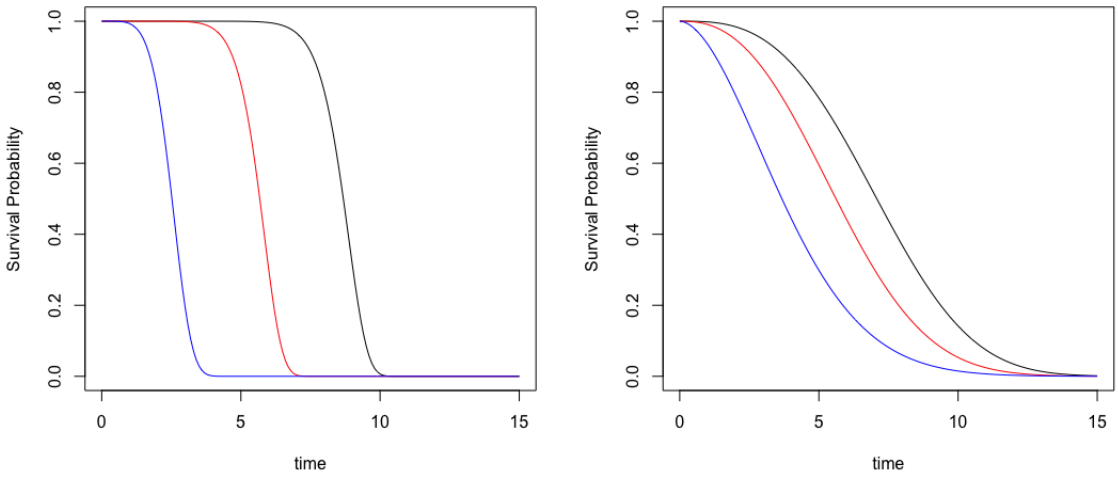

Figure 2.1: Time-to-aGVHD distributions for three risk classes: 1. blue; 2. red; 3. black (left: less overlap of times-to-aGVHD; right: more overlap of times-to-aGVHD)

butions, as shown in the left panel in Figure 2.1, times-to-aGVHD are more distinct between latent classes. For example, at week 4, the majority of aGVHD cases come from latent class 1. On the other hand, when there is more overlap in time-to-aGVHD distributions, as shown in the right panel in Figure 2.1, the aGVHD cases at week 4 come from all three risk classes.

We start with a visualization of the different datasets generated with less or more overlap in times-to-aGVHD. A dataset of 200 patients is generated in each scenario,

and the biomarker observations of these patients are shown in Figure 2.2. According

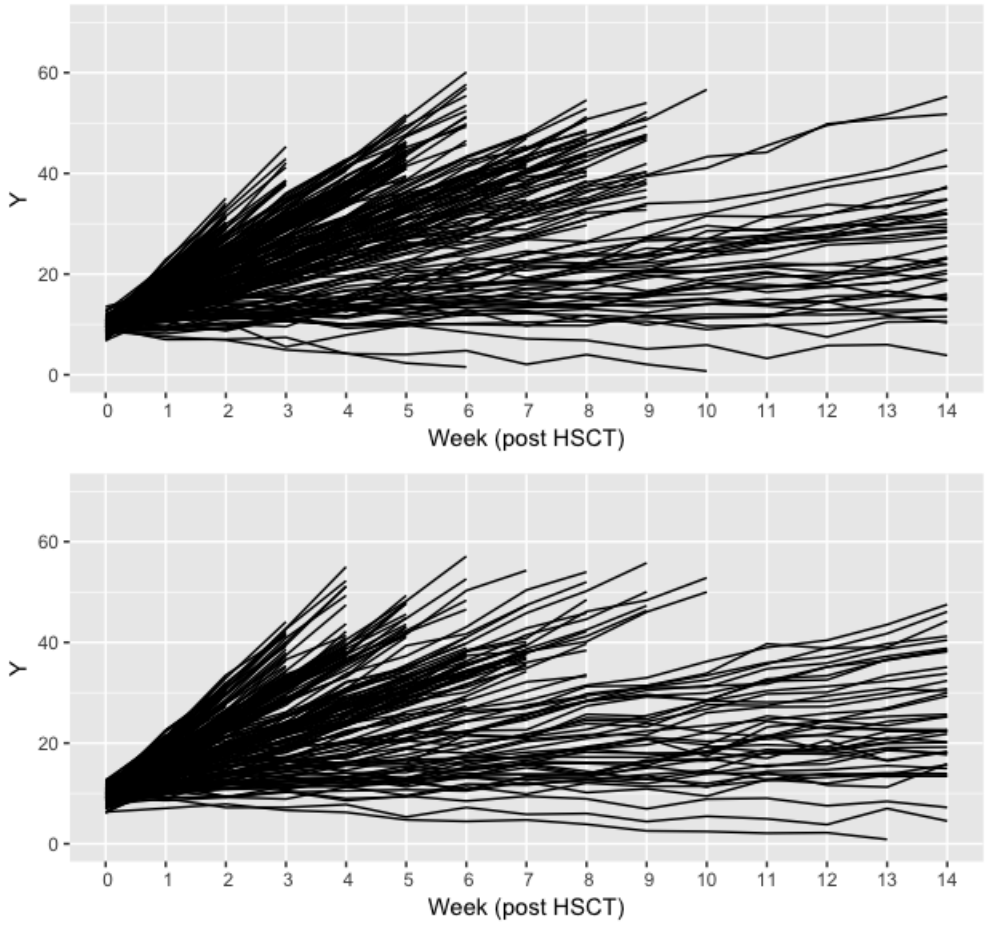

Figure 2.2: Simulated biomarker observations of 200 patients (top: less overlap of time-to-aGVHD distributions; Bottom: more overlap of time-to-aGVHD distributions)

to Figure 2.2, when there is more overlap of times-to-aGVHD between latent classes (shown in the bottom), the distribution of times-to-aGVHD among patients with similar biomarker trajectories is dispersive.

Table 2.1 summarizes the AUCs and BSs of prediction at weeks 1, 2, $\cdots$, 8. Both JMLC and landmark analysis do a better job in distinguishing patients with various risk, and providing a more accurate prediction of aGVHD onset, when there is less overlap of time-to-aGVHD distributions between latent classes in contrast to the

Table 2.1: Mean (SD) of BSs and AUCs of 5,000 simulations with two degrees of overlap in time-to-aGVHD distributions

Prediction made at (week)		1	2	3	4	5	6	7	8
Less overlap of time-to-aGVHD distributions between latent classes									
AUC	JMLC	0.859	0.984	0.918	0.925	0.988	0.907	0.857	0.993
		(0.042)	(0.009)	(0.026)	(0.015)	(0.007)	(0.028)	(0.035)	(0.010)
	LM	0.942	0.971	0.874	0.908	0.968	0.867	0.822	0.973
		(0.019)	(0.022)	(0.037)	(0.022)	(0.033)	(0.053)	(0.046)	(0.018)
BS	JMLC	0.158	0.155	0.073	0.132	0.086	0.114	0.153	0.035
		(0.012)	(0.015)	(0.014)	(0.030)	(0.039)	(0.014)	(0.029)	(0.020)
	LM	0.090	0.062	0.080	0.103	0.063	0.115	0.158	0.059
		(0.020)	(0.026)	(0.015)	(0.016)	(0.039)	(0.030)	(0.021)	(0.022)
More overlap of time-to-aGVHD distributions between latent classes									
AUC	JMLC	0.817	0.923	0.918	0.915	0.895	0.876	0.914	0.971
		(0.064)	(0.019)	(0.017)	(0.025)	(0.025)	(0.031)	(0.031)	(0.020)
	LM	0.871	0.893	0.872	0.873	0.857	0.829	0.879	0.954
		(0.034)	(0.023)	(0.036)	(0.041)	(0.037)	(0.051)	(0.041)	(0.026)
BS	JMLC	0.149	0.156	0.138	0.134	0.135	0.138	0.108	0.059
		(0.018)	(0.031)	(0.027)	(0.026)	(0.017)	(0.017)	(0.026)	(0.025)
	LM	0.103	0.118	0.137	0.137	0.144	0.157	0.130	0.080
		(0.012)	(0.015)	(0.018)	(0.023)	(0.021)	(0.021)	(0.021)	(0.026)

more overlap scenario. This is because in the less overlap scenario, times-to-aGVHD are more distinct between latent classes. Therefore, JMLC, in which a patient's time-to-aGVHD contributes to the latent class identification, results in a better prediction of one's latent class membership, and thus produces a more accurate aGVHD prediction. On the other hand, landmark analysis, in which the latent class membership only depends on the biomarkers, also achieves more accurate aGVHD prediction because the variance of times-to-aGVHD in each latent class is small.

In both the less and more overlap scenarios, JMLC better distinguishes patients who will experience aGVHD in the next two weeks, with higher AUCs starting from week 2. This is because the discrimination ability depends mainly on accurately identifying the latent class membership. JMLC, which incorporates a patient's time-to-aGVHD to the latent class prediction, has more accurate predictions of latent class membership than landmark analysis, thus, it has higher AUCs. However, when we do prediction at week 1, JMLC is inferior to landmark analysis. This is because everyone

is aGVHD-free until week 1, so the fact that a patient is aGVHD-free by week 1 does not contribute to the prediction of latent class membership. Moreover, in contrast to JMLC, landmark analysis uses only the available biomarker observations before the landmark time to fit a model. Thus, it avoids interpreting the noise associated with later biomarkers and has a better AUC than JMLC, when we make a prediction at week 1.

For both simulation scenarios with either less or more overlap of times-to-aGVHD, the prediction accuracy of JMLC is inferior to landmark analysis when only a few biomarker observations are available, reflecting larger BSs. This is because landmark analysis ignores all the biomarker observations and time-to-aGVHD after the landmark time, and thus avoids interpreting the noise associated with future observations. However, when more biomarker observations are collected, JMLC shows better prediction accuracy than landmark analysis in terms of lower BSs. This accuracy gap becomes more obvious when there is more overlap of times-to-aGVHD between latent classes. This is because at each landmark time, we still fix the number of latent classes of aGVHD at four, which, according to Figure 2.2, is not a correct assumption in the later post-HSCT period. For example, after week 7, the majority of patients are from only two latent classes of aGVHD. Assuming the patients are from four latent classes will over-fit the training dataset, and thus result in poor prediction performance.

We also find that landmark analysis is more sensitive than JMLC to the degree of overlap in time-to-aGVHD distributions. This is because landmark analysis builds a model only on the subset of patients who are still at risk at each landmark time. When there is less overlap in times-to-aGVHD between latent classes, the risk set changes dramatically at different landmark times. JMLC, however, fits a model on the whole dataset, incorporating both patients' biomarker history and times-to-aGVHD. Thus,

it is more robust to the change of overlap in time-to-aGVHD distributions.

In this subsection, we have compared the prediction performance of JMLC and landmark analysis, in terms of discrimination and accuracy, under less overlap and strong overlap of times-to-aGVHD scenarios. If we compare the AUCs of JMLC across prediction times, we do not find a clear trend and the AUCs fluctuate. Simply comparing AUCs across time is not recommended because there are different patients at risk at each prediction time. In Table 2.1, we present AUCs calculated with a cumulative sensitivity and a dynamic specificity (Heagerty and Zheng, 2005). The interpretation of cumulative sensitivity is straightforward: that among all the patients at risk at time s, whoever develop the event between $(s, s + t^*)$ are labeled as cases, where s is the prediction time, and t^* is the prediction window. We adopt a dynamic specificity, because the patients in the "control" set at time s might join the "aGVHD-cases" later, so the later the prediction time s is, the fewer the patients in the control set. With a dynamic specificity, the AUC highly depends on the distribution of time-to-aGVHD, and thus it is inappropriate to compare AUCs across time.

As shown in Table 2.1, in the less overlap scenario, the AUC of JMLC at week 2, 5, and 8 are relatively high, compared with other prediction times. This matches what we see in the left panel in Figure 2.1. At week 4, 7, and 10, the majority of aGVHD cases come from one latent class, and nearly all the aGVHD-free patients come from the other latent classes. Therefore, the 2-week predictions at week 2, 5 and 8 achieve higher AUCs. On the other hand, we do not see the same pattern of AUCs of JMLC in the more overlap scenario. This is because at week 4, 7, and 10, the aGVHD cases could be from all the risk latent classes, and the aGVHD-free patients could also be from all latent classes. Thus, we do not recommend comparing AUCs across time.

For the remaining two settings, we will simulate data from the less overlap scenario, and explore the effects of measurement error and model mis-specification.

2.5.2 Effect of Biomarker Measurement Error

Second, we focus on comparing the relative performance of JMLC and landmark under two degrees of measurement error. The explanation of certain patterns in AUCs and BSs that are caused by the less overlap time-to-aGVHD distributions between latent classes are omitted in this subsection.

We randomly generate two datasets of 200 patients with large and small biomarker measurement errors, as demonstrated in Figure 2.3.

Table 2.2 summarizes the AUCs and BSs of aGVHD prediction at week 1, 2, $\cdots$, and 8. When the measure error is small, both JMLC and landmark analysis better distinguish patients in high risk of aGVHD and provide more accurate prediction of aGVHD onset, compared with the scenario with large measurement error. However, this superiority is alleviated when more biomarker observations are available at the prediction time. This is because when more biomarker observations are available, both models have more power to eliminate the effect of the measurement error and fit the model with the true biomarker values. Thus, the size of the measurement error does not affect the results when more biomarker observations are available.

When the measurement error is large, JMLC does better than landmark analysis in identifying patients who will experience aGVHD in two weeks when more than two biomarker observations are available. When the measurement error is small, JMLC also better distinguishes patients who will experience aGVHD in two weeks, with higher AUCs at all prediction times than landmark analysis.

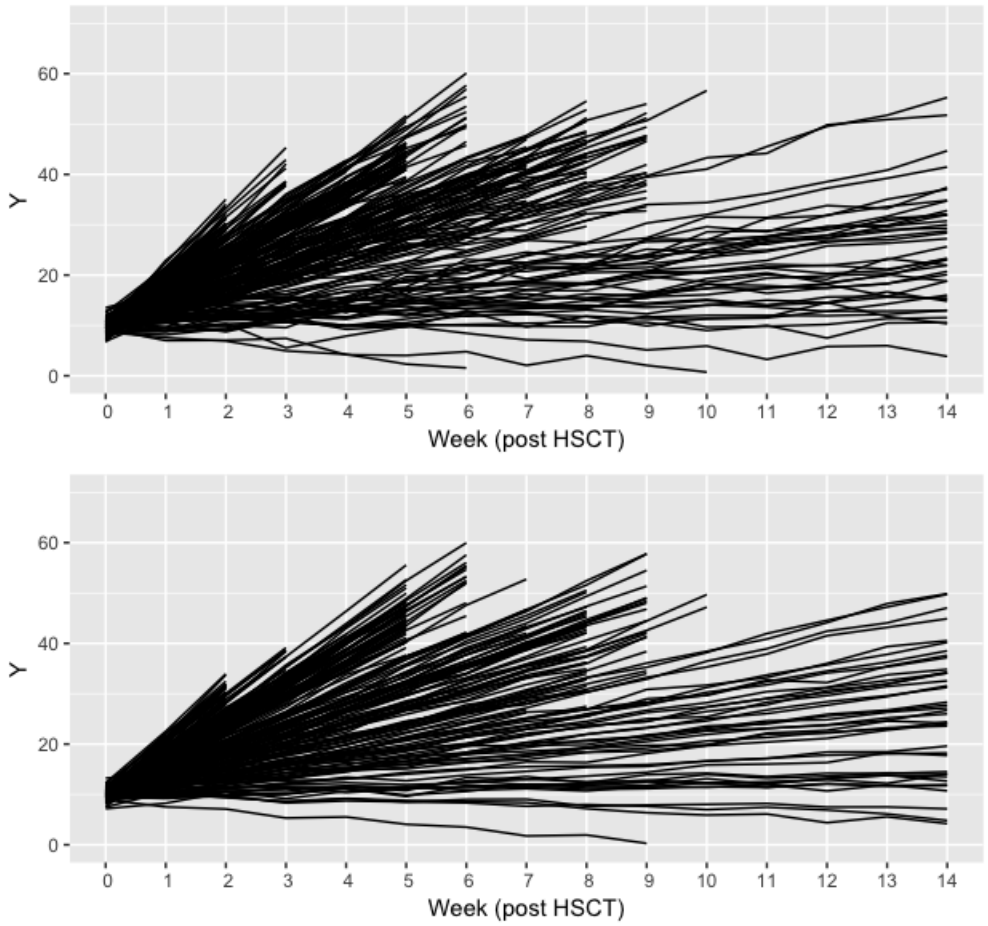

Figure 2.3: Simulated biomarker observations of 200 patients (top: large (sd = 1) measurement error; Bottom: small (sd = 0.5) measurement error)

When the measurement error is large, the absolute prediction accuracy of JMLC is inferior when there are only two or three biomarker observation available, reflecting larger BSs than landmark analysis. When more biomarker observations are collected, JMLC and landmark analysis perform equally well in terms of BSs; and JMLC does perform better than landmark analysis for later predictions when more than eight biomarker observations are available. In the small measurement error scenario, we observe the same trend in BSs as that in the large measurement error scenario.

When we compare between simulations with large and small measurement errors,

Table 2.2: Mean (SD) of BSs and AUCs of 5,000 simulations with two measurement errors of biomarkers

Prediction made at (week)		1	2	3	4	5	6	7	8
large measurement error (sd = 1)									
AUC	JMLC	0.859	0.984	0.918	0.925	0.988	0.907	0.857	0.993
		(0.042)	(0.009)	(0.026)	(0.015)	(0.007)	(0.028)	(0.035)	(0.010)
	LM	0.942	0.971	0.874	0.908	0.968	0.867	0.822	0.973
		(0.019)	(0.022)	(0.037)	(0.022)	(0.033)	(0.053)	(0.046)	(0.018)
BS	JMLC	0.158	0.155	0.073	0.132	0.086	0.114	0.153	0.035
		(0.012)	(0.015)	(0.014)	(0.030)	(0.039)	(0.014)	(0.029)	(0.020)
	LM	0.090	0.062	0.080	0.103	0.063	0.115	0.158	0.059
		(0.020)	(0.026)	(0.015)	(0.016)	(0.039)	(0.030)	(0.021)	(0.022)
small measurement error (sd = 0.5)									
AUC	JMLC	0.967	0.997	0.930	0.932	0.989	0.904	0.853	0.988
		(0.014)	(0.004)	(0.025)	(0.017)	(0.008)	(0.030)	(0.034)	(0.054)
	LM	0.964	0.980	0.884	0.915	0.968	0.866	0.826	0.977
		(0.015)	(0.028)	(0.035)	(0.023)	(0.038)	(0.047)	(0.038)	(0.018)
BS	JMLC	0.154	0.124	0.069	0.103	0.054	0.110	0.155	0.026
		(0.013)	(0.026)	(0.012)	(0.024)	(0.031)	(0.014)	(0.030)	(0.021)
	LM	0.065	0.043	0.077	0.099	0.059	0.110	0.156	0.054
		(0.017)	(0.029)	(0.017)	(0.016)	(0.042)	(0.028)	(0.017)	(0.024)

we find that increasing measurement error in biomarkers weakens the prediction accuracy for both JMLC and landmark analysis, resulting in lower AUCs and larger BSs. However, this negative impact attenuates at later post-HSCT predictions.

2.5.3 Effect of Model Specification

A main feature of landmark analysis is that we avoid specifying any functional form of biomarker trajectories, but assume a multivariate normal distribution of biomarkers. Thus, we want to evaluate the impact of model misspecification on JMLC, especially when JMLC chooses a basic functional form to describe how biomarkers change over time. Figure 2.4 presents one example of simulated patients' dataset, with biomarkers changing non-linearly over time, and the functional forms various across latent classes of aGVHD.

According to the dataset presented in Figure 2.4, assuming that biomarkers change

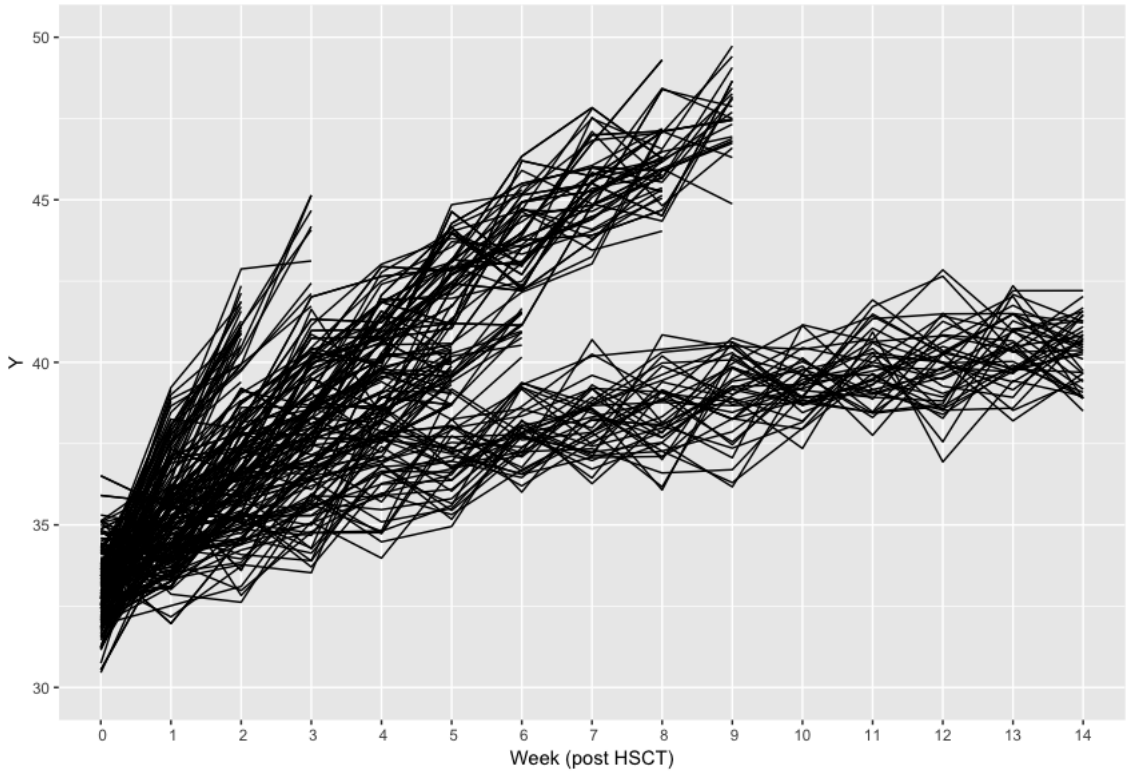

Figure 2.4: Simulated biomarker observations of patients from four latent risk classes of aGVHD, with biomarkers changing unlinearly over time

linearly over time is specious. We fit two separate models to this dataset, one is landmark analysis, and the other is a JMLC with biomarkers changing linearly after BMT. The results of 5,000 simulations are shown in Table 2.3.

When biomarker observations change non-linearly over time, there is a noticeable decrease in the prediction performance of JMLC, reflecting reduced AUCs and inflated BSs. Landmark analysis, on the other hand, provides consistent discrimination ability and accuracy, regardless of how biomarkers change over time.

When biomarker observations change non-linearly over time, JMLC, compared with landmark analysis, better distinguishes patients in high risk of aGVHD when more than three biomarker observations are available. This is because JMLC in-

Table 2.3: Mean (SD) of BSs and AUCs of 5,000 simulations with two functional forms of biomarkers

Prediction made at (week)		1	2	3	4	5	6	7	8
Linear functional forms of biomarkers									
AUC	JMLC	0.859	0.984	0.918	0.925	0.988	0.907	0.857	0.993
		(0.042)	(0.009)	(0.026)	(0.015)	(0.007)	(0.028)	(0.035)	(0.010)
	LM	0.942	0.971	0.874	0.908	0.968	0.867	0.822	0.973
		(0.019)	(0.022)	(0.037)	(0.022)	(0.033)	(0.053)	(0.046)	(0.018)
BS	JMLC	0.158	0.155	0.073	0.132	0.086	0.114	0.153	0.035
		(0.012)	(0.015)	(0.014)	(0.030)	(0.039)	(0.014)	(0.029)	(0.020)
	LM	0.090	0.062	0.080	0.103	0.063	0.115	0.158	0.059
		(0.020)	(0.026)	(0.015)	(0.016)	(0.039)	(0.030)	(0.021)	(0.022)
Non-linear functional forms of biomarkers									
AUC	JMLC	0.523	0.964	0.841	0.814	0.909	0.868	0.849	0.990
		(0.147)	(0.019)	(0.077)	(0.126)	(0.101)	(0.057)	(0.045)	(0.052)
	LM	0.944	0.978	0.692	0.792	0.898	0.820	0.836	0.995
		(0.022)	(0.020)	(0.073)	(0.107)	(0.117)	(0.102)	(0.031)	(0.009)
BS	JMLC	0.193	0.120	0.180	0.554	0.512	0.441	0.180	0.005
		(0.101)	(0.035)	(0.036)	(0.065)	(0.076)	(0.080)	(0.043)	(0.007)
	LM	0.104	0.031	0.084	0.151	0.103	0.121	0.136	0.004
		(0.021)	(0.032)	(0.020)	(0.026)	(0.059)	(0.037)	(0.014)	(0.007)

corporates the aGVHD-free time into the latent class prediction, and thus it better predicts the latent class membership than landmark analysis. This advantage offsets the accuracy loss caused by model misspecification.

2.6 Discussion

In our study, landmark analysis is constructed to use all the biomarker information up to the landmark time, and distinguishes patients of different risk classes of aGVHD. In contrast to landmark analysis in other studies, where only the most recent biomarker observation is used at each landmark time, the landmark analysis we proposed have two benefits. First, we adjust for the measurement error of biomarkers. As shown in Table 2.2, when we make the prediction with at least four biomarker observations, the size of measurement error does not affect the prediction performance of landmark analysis. Second, our landmark analysis adjusts for the heterogeneity among patients. Identifying sub-population in various risks is one of the primary goal in practice, because clinicians could assign appropriate intervention according to

patients' risk status and thus avoid over-treatment.

The JMLC we proposed allows one specific latent class as the "cured" class, and it could be applied in many other medical fields. For a few medical conditions, especially chronic disease such as cancer, there is a nonnegligible "cured" fraction of patients whose pathomechanism is distinct to that of susceptible patients. Thus, these "cured" patients' biomarker profiles are different than those of susceptible patients. Failing to identify this "cured" class and treating these patients as censored will lead to biased results, and might cause over-intervention with patients in little risk.

One primary difference between the two proposed methods is how time-to-aGVHD is used to determine the latent class membership. In JMLC, we model the biomarker process and time-to-aGVHD simultaneously, and use both biomarkers and times-to-aGVHD to estimate the latent class each individual belongs to within the model fitting process. In landmark analysis, we split the model fitting into two steps. In the first step, we identify the class-level pattern of biomarkers and individual-level probabilities of patients belonging to each latent class. In the second step, we estimate the distribution of aGVHD in each latent class with the individual-level probabilities as the weight of a patient belonging to a latent class. JMLC utilizes all the data simultaneously and achieves high efficiency and accuracy when assumptions are met. On the other hand, one major benefit of landmark analysis is that it allows flexible parametrization for the biomarker process. As in our proposed landmark analysis, we do not specify a functional form of how biomarkers change over time, and thus reduce the risk of model misspecification. To reduce the bias of parameter estimation caused by modeling biomarkers and times-to-aGVHD separately, the risk set is updated at each landmark time.

Based on the simulation results, JMLC on average shows better discrimination ability on predictions, and more accurate prediction of probability of aGVHD when there are many biomarker observations available. Landmark analysis, however, presents good discrimination ability and prediction accuracy when there are only limited biomarker observations. When predicting time-to-aGVHD in early post-HSCT period, JMLC borrows biomarker information of patients who haven't experienced aGVHD beyond that time period, which may add more noise rather than increase efficiency to the prediction. On the other hand, when predicting time-to-aGVHD in late post-HSCT period, landmark analysis with fixed number of latent classes overtrains the data and adds more noise to model fitting.

One limitation of this study is that both methods require a pre-specified number of latent classes. We assume the population is a mixture of patients from four aGVHD latent groups. Model selection on number of latent classes can be done based on Bayesian information criterion(BIC) or deviance information criterion(DIC) (Proust-Lima et al., 2014). However, we need to repeat the same model fitting process multiple times to find an ideal number of latent classes, and the selection criteria is based on model fitting performance rather than prediction. In this chapter, we avoid discussion on choosing the best number of latent class, and we will address this issue in Chapter V.

In this study, we did not consider competing risks of aGVHD in patients who received HSCT. As introduced in Chapter 1.1, cancer recurrence, infection, and organ failure are major causes of mortality of AML patients; these conditions, together with death, are competing risks which either hinder the observation of aGVHD or modify the chance that aGVHD occurs. In future work, we could consider scenarios in which competing risks are present and times-to-aGVHD will be dependently cen-

sored. In general, we could modify the survival sub-model in JMLC, and the logistic model in landmark analysis, to incorporate these competing risks. Existing competing risk analysis methods could be used to replace the standard Cox regression model.

We also did not consider incorporating other covariates, such as conditioning regimens, donor type, or previous treatment regimens. Nonetheless, our approach is flexible enough to allow the additional modeling of covariates.

CHAPTER III

Dynamic Prediction of Time-to-acute Graft-versus-Host-Disease with Pattern Mixture Model

3.1 Introduction

In the previous chapter, we modeled the association of longitudinal biomarkers and time-to-aGVHD with both JMLC and landmark analysis. Both of the two methods require a pre-specified number of latent classes. One approach to select the best number of latent classes is by model selection. A series of candidate number of latent classes are chosen, and then a separate model is fitted with each number of latent class. The final model is selected based on information based criteria, such as Akaike information criterion (AIC) or Bayesian information criterion (BIC). However, there are several limitations of this procedure. First it is computationally intensive because a separate model is needed for each candidate number of latent classes. Moreover, AIC and BIC are measures on overall model fitting, favouring smaller residuals in the model while penalizing the number of predictors to avoid overfitting. As a result, AIC and BIC are not necessarily measuring prediction performance. In this chapter, we aim to build a tractable but flexible model predicting the onset time of aGVHD given longitudinal biomarker values. Like the methods in Chapter II , our

new model should be able to reflect the varying risk levels of aGVHD, and update the prediction of time-to-aGVHD when a new biomarker observation is available. In contrast to Chapter II, we attempt to simplify model fitting and avoid the need for a pre-specified number of latent classes.

In this chapter, we predict the time-to-aGVHD with the pattern mixture model. The pattern mixture model has been applied to settings with missing data, and the focus has been on its application to longitudinal data with monotone missingness. The pattern mixture model stratifies the data by patterns of missingness, and then models the differences in the distribution of longitudinal data over these patterns. In our settings of aGVHD data, we propose fitting the pattern mixture model to the longitudinal biomarkers, with patterns depending on the time-to-aGVHD. Considering a fictitious dataset with discrete times-to-aGVHD, i.e., there are only a few unique times-to-aGVHD, it is plausible to form one pattern at each time-to-aGVHD, assuming there are enough data at each unique time-to-aGVHD. In practice, we remeasure the time-to-aGVHD in weeks, and round this time-to-aGVHD to the largest next integer. For example, the time-to-aGVHD of a patient at day 24 will be recoded as week 4. Then we assume patients with the same week of aGVHD share the same biomarker trajectory pattern, and the aGVHD-free patients share another biomarker trajectory pattern.

There are several reasons for why we choose to remeasure time-to-aGVHD in weeks. First, this guarantees enough samples for model fitting in each pattern, resulting in better efficiency for parameter estimation. Second, the precision of prediction in weeks is well-accepted for clinicians. Moreover, weekly biomarker screening is scheduled, so the prediction of aGVHD is updated weekly.

In Chapter II, we aimed to predict the onset of aGVHD in the next two weeks with repeatedly collected biomarker observations. In this chapter, we would like to predict the probability of aGVHD in the next week, the week after that, and so forth. In other words, we achieve the whole distribution of future times-to-aGVHD (measured in weeks). For this purpose, we introduce a new assessment of prediction, the Brier Score (BS), which can evaluate the accuracy of prediction on the whole distribution of future times.

The rest of this chapter is organized as follows. First, we introduce how to do model fitting and make predictions with the pattern mixture model. Next, we describe the two assessments we use to measure the performance of the prediction with the pattern mixture model, followed by simulation results and discussion. A conclusion is drawn at the end.

3.2 Pattern Mixture Model Fitting and Prediction

3.2.1 Notation

Define T_i to be the recorded time-to-aGVHD in weeks for subject i, which is the minimum of the time-to-event for subject i, T_i^*, and last follow up time, S_i. We also define δ_i as the indicator of whether subject i experiences aGVHD ($\delta_i = 1$) or is censored ($\delta_i = 0$). Given the properties of the simulated data, for all patients in the aGVHD-free group, $\delta_i = 0$; and for all patients not in aGVHD-free group, $\delta_i = 1$. We also define $\boldsymbol{Y}_i = (Y_i(t_1), Y_i(t_2), \cdots, Y_i(t_{n_i}))$ as the biomarker history of subject i at time $(t_1, t_2, \cdots, t_{n_i})$, where n_i is the total number of biomarker observations for subject i, for $i = 1, 2, \cdots, n$.

For simplicity, we assume that the continuous biomarker observations change lin-

early over time, but it is straightforward to generalize them as other functional forms of time. Random effects $\boldsymbol{b_i} = (b_{0i}, b_{1i})$ are introduced to reflect the individual deviation of the biomarker trajectories from the population mean. Thus, we assume the biomarker observations, $\boldsymbol{Y_i}$, follow a multivariate Gaussian distribution with pattern-specified mean profile:

$$\boldsymbol{Y_i}|T_i \;=\; \boldsymbol{X_i}\boldsymbol{\beta_{T_i}} + \boldsymbol{Z_i}\boldsymbol{b_i} + \boldsymbol{\epsilon_i} \tag{3.1}$$

where $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ are the design matrices for subject i with the first column all 1s and the second column the biomarker observation times, $\boldsymbol{b_i}$ are random effects with variance-covariance matrix D, $\boldsymbol{\epsilon_i}$ is the the measurement error which follows a Gaussian distribution with mean $\mathbf{0}$ and variance $\sigma^2\boldsymbol{I}$, and $\boldsymbol{\beta_{T_i}} = (\beta_{0,T_i}, \beta_{1,T_i})'$ is the fixed effects associated with the mean pattern profile. Define $\boldsymbol{\xi} = (\boldsymbol{\beta_T} = (\boldsymbol{\beta_{T_1}}, \boldsymbol{\beta_{T_2}}, \cdots, \boldsymbol{\beta_{T_J}})', D, \sigma^2)$ as the parameters of interest, where J represents the number of unique values of times-to-aGVHD. By fitting a linear mixed model with time-to-aGVHD as a predictor, we obtain the MLE of $\boldsymbol{\xi}$.

3.2.2 Prediction

The marginal distribution of observed biomarkers is a finite mixture of Gaussian distributions, and the posterior probability of a new patient m developing aGVHD at week j, with the observed biomarkers history by week k, $\boldsymbol{Y_m} = (Y_{m1}, Y_{m2}, \cdots, Y_{m,k})$, is:

$$Pr(T_m = j|\boldsymbol{Y_m}, \boldsymbol{\xi}) \;=\; \frac{f(\boldsymbol{Y_m}|\boldsymbol{\beta_{T_j}}, D, \sigma^2)Pr(T_m = j)}{\sum_{l=k}^{J} f(\boldsymbol{Y_m}|\boldsymbol{\beta_{T_l}}, D, \sigma^2)P(T_m = l)} \tag{3.2}$$

for $j \geq k$. The biomarker observations $\boldsymbol{Y_m}$ given time-to-aGVHD, T_j, follow a multivariate normal distribution with mean $\boldsymbol{X_m}\boldsymbol{\beta_{T_j}}$ and variance-covariance matrix

$\boldsymbol{Z_m D Z'_m} + \sigma^2 \boldsymbol{I}$.

3.3 Predictive Accuracy Measures for Dynamic Predictions

Our first measure of predictive accuracy is the Brier Score (BS), which measures the calibration of probabilistic predictions (Brier, 1950). In a multi-class setting, it is defined as $BS = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J}(f_{ij} - o_{ij})^2$, where f_{ij} is the forecasted probability of subject i ($i = 1, 2, \cdots, n$) belonging to the category j ($j = 1, 2, \cdots, J$), and o_{ij} is the actual outcome of subject i (1 if subject i belongs to category j, 0 otherwise). In a case when there are 11 categories, the BS of a random guess is 10/11, and a perfect prediction would achieve BS at 0. To make the BS comparable to another metric we present later, we will subtract the BS from 1, so that a perfect prediction has BS = 1; a random prediction has BS close to 0.

In our study, we propose a dynamic BS, which is defined as:

$$BS(k) = 1 - \frac{1}{\sum_{i=1}^{n} I(T_i \geq k)} \sum_{i=1}^{n}\sum_{j=1}^{J} I(T_i \geq k)[f_{ij}(\boldsymbol{Y_i}) - o_{ij}]^2 \tag{3.3}$$

which is the sum of squared prediction errors across all subjects who are still at risk at the prediction time k. Note that the BS can be written as:

$$\begin{aligned} BS(k) &= 1 - \frac{1}{\sum_{i=1}^{n} I(T_i \geq k)} \sum_{i=1}^{n}\sum_{j=1}^{J} I(T_i \geq k)([f_{ij}(\boldsymbol{Y_i}) - E(o_{ij})]^2 + [E(o_{ij}) - o_{ij}]^2) \\ &= \text{Bias}^2[f_{ij}] + \text{Var}[f_{ij}] \end{aligned} \tag{3.4}$$

Therefore, BS summarizes both the accuracy and uncertainty of the prediction.

Our second measure of the prediction accuracy is the dynamic prediction accuracy rate (PAR). We define PAR at time k to be the proportion of accurate prediction

given biomarker observations up to time k in patients who are at risk at time k, i.e.

$$PAR(k) = \frac{\sum_{i=1}^{n}\sum_{t=k}^{J} I(T_i = t)I(P(T_i = t) \geq P(T_i = l) \text{ for any } l \neq t)}{\sum_{i=1}^{n} I(T_i \geq k)} \quad (3.5)$$

PAR will increase if we widen the windows of accuracy of prediction, perhaps by including predictions that are one week earlier or later than the actual event time only.

3.4 Simulation and Result

In this section we will simulate data under three settings: (1) a joint model with eleven latent classes, (2) a joint model with four latent classes, and (3) a joint model with four latent class and shared random effects. These three settings reflect different fundamental assumptions on the relationship between time-to-aGVHD, biomarkers and latent classes. Setting (1) assumes the latent class can be almost defined by time-to-aGVHD, and vice versa. This implies that there are less overlappping in the distributions of time-to-aGVHD of each pattern. Setting (2) adopts the same assumptions as setting (1) by assuming the patients are of varying latent classes of aGVHD, and patients within one latent class share the same distribution of time-to-aGVHD. However,unlike setting (1), in setting (2) patients in one latent class share a more disperse distribution of time-to-aGVHD and biomarker trajectories. Patients have similar biomarker trajectories may end up with obvious different times-to-aGVHD, so it is hard to distinguish the latent class from each other just based on times-to-aGVHD, compared with setting (1). Setting (3) considers that individual deviations of biomarker trajectories also contribute to the variation of time-to-aGVHD within each latent class. In other words, the fundamental assumption of conditional independence of JMLC is violated, and both the latent classes and these individual deviations contribute to the correlation between biomarkers and times-to-aGVHD.

In all three simulation settings, biomarker screening is scheduled right after HSCT (baseline) and weekly thereafter until the onset of aGVHD. An uninformative baseline biomarker level is assumed, so at least two biomarker observations are needed to make a prediction for the onset of aGVHD. For simplicity, we specify that the biomarkers change linearly over time with measurement errors. Patients within one latent class share the same mean intercept and slope of biomarkers, with individual deviations of intercepts and slopes. In each simulation setting, we explore the performance of our model under seven scenarios defined by the variance-covariance structure of individual deviations of the intercept and the slope, as well as the variance of the measurement error. The details of seven scenarios can be found in Table 3.1. Compared with Scenario 1-3, biomarkers generated under Scenario 4-6 are more distinct between patterns, given the smaller variance of deviation of biomarker trajectory. In contrast with Scenario 4, Scenario 7 has a larger measurement error of biomarkers.

Table 3.1: Simulation scenarios with various covariance of random effects and variance of measurement error

Scenario	$Var(b_{0i})$	$\mathrm{Var}(b_{1i})$	$\rho(b_{0i}, b_{1i})$	$sd(\epsilon)$
1	0.16	0.16	0	0.5
2	0.16	0.16	-0.5	0.5
3	0.16	0.16	0.5	0.5
4	0.0625	0.0625	0	0.5
5	0.0625	0.0625	-0.5	0.5
6	0.0625	0.0625	0.5	0.5
7	0.0625	0.0625	0	1

A series of prediction times are set weekly from week one until week ten, right after biomarker screening. We run 1,000 simulations and compute the means and standard deviations of the resulting BS and PAR at each future time s, where $s = 1, 2, \cdots, 10th$ week. In each simulation setting, we compare the prediction accuracy of our model with the theoretically best prediction that could be achieved. The latter is calculated based on the "true" value of parameters, and reflects the variation of prediction even

when parameters of interest are correctly estimated. Through these comparisons, we are able to quantify the loss of prediction accuracy caused by assigning latent class membership solely on a patient's time-to-aGVHD.

3.4.1 Simulation from Joint Model with Eleven Latent Classes

We start with simulation setting (1), where we simulate data from a joint model with eleven latent classes. We choose eleven latent classes because we will group the patients into eleven patterns according to the week they develop aGVHD. We assume that the patients of each latent class share the same time-to-aGVHD distribution, which is defined as a Weibull distribution. Two hundred patients are simulated as the training dataset, and another 200 patients from same population are simulated as the test dataset. In this population, around 5% subjects would most likely experience aGVHD one week after BMT, another 15% might undergo aGVHD onset two weeks after BMT, and so forth. The details of simulation parameters can be found in Table 3.2.

Table 3.2: Simulation parameters for joint modeling with eleven latent classes

Latent	Population	Time to aGVHD (Weibull)		Biomarker Trajectory	
Class	Proportion	λ	κ	Intercept	Slope
1	5%	11.605	34	10	0
2	5%	10.604	31	10	1
3	5%	9.603	28	10	2
4	5%	8.601	25	10	3
5	5%	7.599	22	10	4
6	10%	6.596	19	10	5
7	10%	5.592	16	10	6
8	15%	4.594	14	10	7
9	20%	3.588	11	10	8
10	15%	2.577	8	10	9
11	5%	1.55	5	10	10

Although in this setting, the distributions of time-to-aGVHD in each pattern are

concentrated, the simulated times-to-aGVHD can still take various values. Thus, analyzing the data simulated from setting (1) with the pattern mixture model, we might group patients from different latent classes into one pattern, so the biomarker trajectory estimation of one latent class might be biased. This reflects the bias portion of BS as shown in Equation 3.4. In other words, when using the pattern mixture model to analyze data simulated from a joint model, BS contains bias due to grouping patients based on the time-to-aGVHD only.

In order to quantify the accuracy loss of prediction by this bias, we calculate the "true" probability of aGVHD under joint modeling with eleven latent classes as:

$$\begin{aligned}
&P(k+l \leq T_m \leq k+l+1 | T_m \geq k, \boldsymbol{Y}_m, \boldsymbol{\theta}) \\
&= \sum_{j=1}^{J} P(k+l \leq T_m \leq k+l+1 | T_m \geq k, c_m = j, \boldsymbol{\theta}) P(c_m = j | T_m \geq k, \boldsymbol{Y}_m, \boldsymbol{\theta}) \\
&= \frac{\sum_{j=1}^{J} Pr(c_m = j) f(\boldsymbol{Y}_m | c_m = j, \boldsymbol{\theta}) [S(k+l | c_m = j, \boldsymbol{\theta}) - S(k+l+1 | c_m = j, \boldsymbol{\theta})]}{\sum_{j=1}^{J} Pr(c_m = j) f(\boldsymbol{Y}_m | c_m = j, \boldsymbol{\theta}) S(k | c_m = j, \boldsymbol{\theta})}
\end{aligned} \tag{3.6}$$

where $\boldsymbol{\theta}$ is the simulation parameters as listed in Table 3.2, $\boldsymbol{Y}_m$ is the biomarker observations history of subject m, and k is the prediction time.

Table 3.3 summarizes the mean BSs of the pattern mixture model, the "true" model, and the mean and standard deviation of their difference. In each scenario, the pattern mixture model has lower BSs relative to the "true" model. However, compared with their standard deviations, these losses are not significantly different from 0. Moreover, compared with the BS of the "true" model, pattern mixture model losses approximate 5% of the BS.

Table 3.3: Brier Score of the pattern mixture model under simulation setting (1)

Prediction at week	1	2	3	4	5	6	7	8	9	10
Scenario 1										
True	0.384	0.511	0.571	0.585	0.582	0.607	0.597	0.614	0.653	0.726
PM	0.365	0.484	0.540	0.549	0.539	0.563	0.547	0.564	0.603	0.677
Difference (mean)	0.019	0.027	0.032	0.036	0.042	0.044	0.050	0.050	0.050	0.049
Difference (SD)	0.012	0.015	0.018	0.023	0.029	0.034	0.043	0.049	0.056	0.067
Scenario 2										
True	0.415	0.563	0.618	0.626	0.622	0.644	0.627	0.644	0.676	0.750
PM	0.395	0.530	0.581	0.585	0.574	0.594	0.572	0.589	0.623	0.696
Difference (mean)	0.021	0.033	0.037	0.041	0.047	0.050	0.055	0.055	0.053	0.053
Difference (SD)	0.012	0.018	0.021	0.026	0.032	0.037	0.045	0.050	0.057	0.072
Scenario 3										
True	0.365	0.490	0.558	0.580	0.584	0.615	0.603	0.627	0.663	0.739
PM	0.346	0.465	0.528	0.543	0.540	0.568	0.550	0.572	0.613	0.690
Difference (mean)	0.019	0.025	0.031	0.037	0.044	0.047	0.053	0.055	0.051	0.050
Difference (SD)	0.012	0.015	0.018	0.023	0.029	0.034	0.043	0.048	0.056	0.069
Scenario 4										
True	0.427	0.604	0.673	0.688	0.686	0.707	0.697	0.715	0.744	0.804
PM	0.407	0.567	0.630	0.640	0.634	0.653	0.638	0.656	0.690	0.753
Difference (mean)	0.020	0.037	0.044	0.047	0.052	0.054	0.060	0.059	0.054	0.051
Difference (SD)	0.013	0.020	0.022	0.027	0.032	0.037	0.046	0.050	0.055	0.065
Scenario 5										
True	0.450	0.645	0.704	0.709	0.705	0.723	0.712	0.725	0.754	0.812
PM	0.430	0.603	0.659	0.662	0.653	0.670	0.654	0.667	0.698	0.762
Difference (mean)	0.021	0.042	0.045	0.047	0.052	0.053	0.058	0.057	0.056	0.051
Difference (SD)	0.013	0.020	0.022	0.026	0.031	0.036	0.043	0.048	0.056	0.061
Scenario 6										
True	0.412	0.579	0.657	0.679	0.683	0.705	0.697	0.713	0.744	0.803
PM	0.394	0.547	0.617	0.634	0.632	0.654	0.640	0.656	0.690	0.756
Difference (mean)	0.018	0.032	0.040	0.045	0.050	0.051	0.057	0.057	0.054	0.047
Difference (SD)	0.012	0.017	0.021	0.025	0.031	0.035	0.042	0.049	0.056	0.063
Scenario 7										
True	0.297	0.487	0.612	0.657	0.671	0.702	0.697	0.715	0.744	0.807
PM	0.284	0.465	0.579	0.614	0.620	0.648	0.637	0.655	0.688	0.756
Difference (mean)	0.013	0.021	0.034	0.043	0.051	0.055	0.060	0.060	0.056	0.051
Difference (SD)	0.009	0.014	0.020	0.025	0.031	0.036	0.044	0.050	0.053	0.061

The BS values for Scenarios 4-6 are larger than those for Scenarios 1-3. This is because the individual deviations of biomarker trajectories in Scenarios 4-6 have smaller variance, leading to larger between-latent class variance relative to within-latent class variance. We also note that when the individual intercepts and slopes are negatively correlated, the BS values are larger compared to scenarios with independent or positively correlated intercepts and slopes. Moreover, a larger measurement error (Scenario 7) leads to poorer BS values when available biomarker information is insufficient, especially when making prediction with no more than four repeated

biomarker observations. However, this influence is reduced when more biomarker observations are available.

Table 3.4 lists the mean PARs from "true" joint modeling, mean from of pattern mixture model, and their mean differences and standard deviation. From Table 3.4 we detect the loss of PAR of pattern mixture model compared with the "true" mode. Similar to the BS, compared with the standard deviation, these PAR losses are not significantly away from 0. Moreover, compared with the size of PAR of the "true" model, the pattern mixture model losses approximate 5% of the PAR.

We can find the same pattern of PAR loss in various scenarios as of BS loss. Among various scenarios, Scenario 4-6 show larger PARs, compared with Scenario 1-3. When the individual intercept and slope are negatively correlated, the PAR achieves larger values compared with scenarios having independent or positively correlated intercepts and slopes. Moreover, large measurement error (Scenario 7) do worsen PARs when available biomarker information is insufficient, but this influence is diluted when more than four biomarker observations are available. In contrast to the BS loss, the PAR loss is easy to understand and interpret. For example, under Scenario 6 and at prediction time 4, the pattern mixture model incorrectly predicts three out of 100 patients.

3.4.2 Simulation from Joint Model with Four Latent Classes

Now we assume the patients are from four latent classes of aGVHD. Compared with Section 3.4.1, the number of latent classes is far smaller than the number of patterns we choose in the pattern mixture model. Similar to Section 3.4.1, patients in the same latent class share the same biomarker mean trajectory with individual deviations, and same aGVHD probability, which follows a Weibull distribution. In

Table 3.4: PAR of the pattern mixture model under simulation setting (1)

Predictionatweek	1	2	3	4	5	6	7	8	9	10
Scenario 1										
True	0.505	0.638	0.693	0.705	0.704	0.724	0.718	0.732	0.762	0.816
PM	0.483	0.612	0.663	0.672	0.667	0.685	0.673	0.687	0.719	0.778
Difference (mean)	0.022	0.026	0.029	0.034	0.038	0.039	0.045	0.045	0.042	0.038
Difference (SD)	0.028	0.027	0.029	0.036	0.043	0.048	0.058	0.064	0.072	0.086
Scenario 2										
True	0.543	0.691	0.738	0.745	0.743	0.759	0.748	0.761	0.783	0.838
PM	0.519	0.662	0.706	0.710	0.703	0.715	0.699	0.713	0.739	0.796
Difference (mean)	0.024	0.029	0.032	0.035	0.041	0.044	0.049	0.048	0.044	0.042
Difference (SD)	0.027	0.027	0.030	0.036	0.043	0.047	0.058	0.064	0.070	0.081
Scenario 3										
True	0.483	0.617	0.680	0.701	0.708	0.732	0.723	0.742	0.772	0.828
PM	0.460	0.590	0.649	0.666	0.667	0.688	0.674	0.694	0.727	0.790
Difference (mean)	0.022	0.027	0.031	0.035	0.041	0.044	0.049	0.048	0.045	0.038
Difference (SD)	0.028	0.028	0.029	0.036	0.042	0.048	0.060	0.066	0.072	0.083
Scenario 4										
True	0.557	0.730	0.789	0.802	0.803	0.817	0.811	0.824	0.843	0.884
PM	0.534	0.700	0.754	0.764	0.763	0.775	0.765	0.778	0.802	0.848
Difference (mean)	0.023	0.031	0.035	0.037	0.040	0.043	0.047	0.045	0.041	0.036
Difference (SD)	0.027	0.027	0.028	0.033	0.036	0.042	0.051	0.055	0.060	0.067
Scenario 5										
True	0.582	0.769	0.816	0.821	0.821	0.832	0.825	0.834	0.853	0.890
PM	0.559	0.738	0.784	0.787	0.784	0.792	0.782	0.791	0.812	0.860
Difference (mean)	0.023	0.031	0.033	0.035	0.037	0.040	0.043	0.042	0.040	0.030
Difference (SD)	0.026	0.025	0.026	0.03	0.033	0.039	0.046	0.051	0.060	0.059
Scenario 6										
True	0.538	0.708	0.774	0.794	0.800	0.815	0.811	0.823	0.843	0.882
PM	0.517	0.679	0.743	0.758	0.761	0.774	0.767	0.778	0.802	0.853
Difference (mean)	0.021	0.028	0.031	0.036	0.039	0.041	0.045	0.045	0.041	0.029
Difference (SD)	0.027	0.025	0.027	0.031	0.035	0.038	0.047	0.051	0.059	0.061
Scenario 7										
True	0.399	0.613	0.733	0.775	0.790	0.814	0.811	0.824	0.843	0.885
PM	0.381	0.591	0.703	0.739	0.750	0.770	0.765	0.778	0.803	0.854
Difference (mean)	0.017	0.022	0.030	0.036	0.040	0.044	0.047	0.046	0.040	0.031
Difference (SD)	0.029	0.027	0.028	0.031	0.037	0.042	0.051	0.057	0.060	0.065

contrast to Section 3.4.1, in Setting (2) patients in the same latent class share a more widespread distribution of time-to-aGVHD, so the variation of prediction increases. The details of simulation parameters can be found in Table 3.5.

As in Section 3.4.1, we calculate the means BS and PAR values from the pattern mixture model and its theoretically best counterpart, together with the difference between the two. The BSs and PARs are shown in Table 3.6 and Table 3.7 respectively.

Table 3.5: Simulation parameters for joint modeling with four latent classes

Latent Class	Population Proportion	Time to aGVHD (Weibull)		Biomarker Trajectory	
		λ	κ	Intercept	Slope
1	20%	11.256	22	10	1
2	20%	8.91	14	10	4
3	30%	5.89	10	10	7
4	30%	2.715	5	10	9

The results list in Table 3.6 show a pattern similar to that in Table 3.3. There are consistent losses of BSs from the pattern mixture model, compared with BSs from a "true" model. However, these differences are not significantly different from 0. When we compare the BSs across various scenarios, we also find the same pattern as in Table 3.3. Scenario 4-6 show larger BSs, compared with scenario 1-3. When the individual intercepts and slopes are negatively correlated, the BS achieves larger values compared with scenarios having independent or positively correlated intercepts and slopes. Moreover, large measurement errors lower BSs when available biomarker information is insufficient, especially when making prediction with no more than four repeated biomarker observations.

In contrast to Table 3.3, the BS values in Table 3.6 are relatively smaller. This is because the data simulated from a joint model with four latent classes has wider distribution of times-to-aGVHD, leading to larger variation in prediction, and thus smaller BS values.

Based on Table 3.7, PARs show similar pattern as in Table 3.4, with a reduction in their corresponding values. The maximum PAR we can theoretically achieve is around 60%, which means we can only accurately predict the times-to-aGVHD for 60 out of 100 patients. In this case, it would be helpful if we can widen the window of accuracy of PARs, including predictions that are one week earlier or later than the exact time-to-aGVHD.

Table 3.6: Brier Score of the pattern mixture model under simulation setting (2)

Predictionatweek	1	2	3	4	5	6	7	8	9	10
Scenario 1										
True	0.317	0.389	0.380	0.360	0.406	0.452	0.384	0.441	0.529	0.517
PM	0.299	0.364	0.352	0.330	0.375	0.420	0.345	0.403	0.498	0.482
Difference (mean)	0.018	0.025	0.029	0.031	0.031	0.032	0.039	0.038	0.031	0.035
Difference (SD)	0.013	0.015	0.018	0.020	0.022	0.023	0.030	0.034	0.034	0.048
Scenario 2										
True	0.346	0.417	0.405	0.377	0.424	0.474	0.396	0.455	0.547	0.516
PM	0.324	0.389	0.373	0.344	0.391	0.440	0.355	0.414	0.515	0.480
Difference (mean)	0.022	0.028	0.031	0.033	0.034	0.034	0.042	0.041	0.032	0.036
Difference (SD)	0.014	0.016	0.019	0.021	0.022	0.026	0.033	0.037	0.034	0.049
Scenario 3										
True	0.308	0.389	0.387	0.368	0.416	0.468	0.392	0.451	0.544	0.518
PM	0.295	0.368	0.360	0.338	0.385	0.437	0.353	0.413	0.514	0.484
Difference (mean)	0.014	0.021	0.027	0.030	0.031	0.031	0.040	0.038	0.030	0.034
Difference (SD)	0.013	0.015	0.017	0.019	0.021	0.023	0.030	0.033	0.034	0.049
Scenario 4										
True	0.361	0.441	0.432	0.398	0.448	0.501	0.414	0.474	0.571	0.519
PM	0.341	0.413	0.401	0.364	0.413	0.467	0.371	0.431	0.542	0.484
Difference (mean)	0.020	0.028	0.031	0.034	0.035	0.034	0.043	0.043	0.029	0.035
Difference (SD)	0.014	0.016	0.019	0.021	0.023	0.026	0.033	0.037	0.033	0.047
Scenario 5										
True	0.381	0.453	0.439	0.402	0.451	0.509	0.418	0.477	0.576	0.519
PM	0.357	0.422	0.406	0.366	0.415	0.472	0.372	0.433	0.549	0.488
Difference (mean)	0.024	0.031	0.033	0.036	0.036	0.036	0.046	0.044	0.026	0.032
Difference (SD)	0.014	0.017	0.020	0.022	0.024	0.026	0.033	0.038	0.033	0.045
Scenario 6										
True	0.349	0.437	0.431	0.399	0.449	0.506	0.416	0.477	0.574	0.520
PM	0.332	0.411	0.402	0.366	0.415	0.472	0.372	0.435	0.545	0.483
Difference (mean)	0.017	0.026	0.029	0.033	0.035	0.034	0.044	0.042	0.029	0.036
Difference (SD)	0.014	0.016	0.019	0.021	0.024	0.026	0.032	0.038	0.034	0.050
Scenario 7										
True	0.293	0.419	0.425	0.396	0.446	0.502	0.413	0.473	0.572	0.518
PM	0.276	0.394	0.396	0.364	0.413	0.468	0.370	0.432	0.543	0.484
Difference (mean)	0.017	0.025	0.029	0.032	0.034	0.034	0.043	0.042	0.029	0.034
Difference (SD)	0.011	0.015	0.019	0.021	0.024	0.026	0.033	0.037	0.035	0.049

3.4.3 Joint Model with Shared Random Effects and Four Latent Classes

Same as JMLC, another counterpart of the pattern mixture model, JMSR also links longitudinal data to a primary event. However, JMSR assumes the patients have the same pattern of biomarker trajectories over time, and their times-to-aGVHD depend on only the individual deviation of biomarkers from the global mean. Here we generalize the assumption of both JMLC and JMSR, assuming that patients are inherent of various risk groups of aGVHD, and their times-to-aGVHD depend on both

Table 3.7: PAR of the pattern mixture model under simulation setting (2)

Predictionatweek	1	2	3	4	5	6	7	8	9	10
Scenario 1										
True	0.459	0.520	0.503	0.485	0.521	0.557	0.496	0.541	0.618	0.590
PM	0.436	0.492	0.468	0.448	0.481	0.520	0.451	0.497	0.584	0.546
Difference (mean)	0.022	0.028	0.035	0.038	0.040	0.037	0.045	0.045	0.034	0.044
Difference (SD)	0.028	0.030	0.037	0.042	0.044	0.047	0.059	0.062	0.065	0.090
Scenario 2										
True	0.487	0.545	0.523	0.499	0.535	0.573	0.505	0.551	0.629	0.585
PM	0.461	0.515	0.487	0.459	0.495	0.534	0.457	0.503	0.595	0.543
Difference (mean)	0.026	0.030	0.036	0.040	0.041	0.039	0.048	0.048	0.033	0.043
Difference (SD)	0.029	0.031	0.038	0.042	0.043	0.049	0.061	0.065	0.064	0.086
Scenario 3										
True	0.450	0.521	0.508	0.491	0.529	0.570	0.503	0.550	0.631	0.592
PM	0.432	0.495	0.475	0.454	0.491	0.532	0.456	0.504	0.597	0.547
Difference (mean)	0.018	0.025	0.033	0.037	0.038	0.038	0.047	0.046	0.034	0.045
Difference (SD)	0.028	0.030	0.038	0.041	0.043	0.050	0.060	0.064	0.068	0.094
Scenario 4										
True	0.502	0.567	0.547	0.519	0.557	0.596	0.522	0.569	0.650	0.591
PM	0.479	0.539	0.513	0.483	0.519	0.561	0.477	0.525	0.618	0.547
Difference (mean)	0.023	0.028	0.034	0.037	0.038	0.035	0.044	0.043	0.032	0.044
Difference (SD)	0.028	0.030	0.037	0.040	0.042	0.047	0.058	0.062	0.066	0.094
Scenario 5										
True	0.520	0.576	0.552	0.522	0.559	0.602	0.525	0.571	0.654	0.593
PM	0.493	0.547	0.519	0.484	0.521	0.565	0.478	0.527	0.624	0.551
Difference (mean)	0.027	0.029	0.034	0.037	0.038	0.037	0.047	0.044	0.029	0.042
Difference (SD)	0.028	0.030	0.036	0.039	0.041	0.046	0.057	0.060	0.061	0.086
Scenario 6										
True	0.491	0.563	0.547	0.520	0.558	0.601	0.524	0.572	0.653	0.593
PM	0.470	0.536	0.514	0.483	0.519	0.564	0.478	0.527	0.620	0.546
Difference (mean)	0.021	0.027	0.033	0.036	0.039	0.036	0.047	0.044	0.033	0.047
Difference (SD)	0.028	0.031	0.037	0.041	0.043	0.047	0.058	0.063	0.064	0.090
Scenario 7										
True	0.431	0.546	0.541	0.516	0.555	0.595	0.519	0.566	0.650	0.589
PM	0.412	0.520	0.508	0.480	0.516	0.559	0.473	0.523	0.619	0.546
Difference (mean)	0.020	0.026	0.033	0.036	0.039	0.036	0.046	0.043	0.031	0.043
Difference (SD)	0.027	0.030	0.039	0.042	0.044	0.048	0.060	0.063	0.067	0.093

the risk group and their individual deviations of biomarker trajectory.

In this setting, the biomarker observations $\boldsymbol{Y_i}$ and time-to-aGVHD T_i are assumed independent conditioning on random effects $\boldsymbol{b_i}$, which follow a mixture of multivariate normal distribution. Recently there are some researchers trying to extend joint modeling to incorporate mixture distribution of shared random effect with simplified binary outcomes. Given the increasing numbers of parameters, and the booming difficulty in constructing the likelihood, these researchers adopted the Bayesian framework to

achieve posterior predictive distribution on outcomes (Jiang et al., 2015). This approach can thoroughly remove bias, but it is computationally intense.

Here, we use the pattern mixture model to avoid the technique difficulties raised by shared random effects of mixture distributions. Its performance is evaluated by comparing the prediction with the ideal counterpart when all the parameters are known ahead. Similar to Equation 3.6, the "true" probability of aGVHD under setting (3) can be approximated by the following:

$$
\begin{aligned}
&Pr(k+l \leq T_m \leq k+l+1 | T_i \geq k, \boldsymbol{Y}_m, \boldsymbol{\theta}) \\
&= \int Pr(k+l \leq T_m \leq k+l+1 | T_i \geq k, \boldsymbol{Y}_m, \boldsymbol{b_m}, \boldsymbol{\theta}) Pr(\boldsymbol{b_m} | T_m \geq k, \boldsymbol{Y}_m, \boldsymbol{\theta}) d\boldsymbol{b_m} \\
&\approx \frac{\sum_{g=1}^{G} Pr(T_m \in (k+l, k+l+1) | T_m > k, \boldsymbol{b_m}^{(g)}, \boldsymbol{\theta}) Pr(T_m \geq k | \boldsymbol{b_m}^{(g)}, \boldsymbol{\theta}) Pr(\boldsymbol{Y}_i | \boldsymbol{b_m}^{(g)}, \theta) Pr(\boldsymbol{b_m}^{(g)} | \boldsymbol{\theta})}{\mathrm{G} \int Pr(T_m \geq k | \boldsymbol{b_m}, \boldsymbol{\theta}) Pr(\boldsymbol{Y}_m | \boldsymbol{b_m}, \boldsymbol{\theta}) Pr(\boldsymbol{b_m} | \boldsymbol{\theta}) d\boldsymbol{b_m}} \\
&= \frac{\sum_{g=1}^{G} Pr(k+l \leq T_m \leq k+l+1 | \boldsymbol{b_m}^{(g)}, \boldsymbol{\theta}) Pr(\boldsymbol{Y}_m | \boldsymbol{b_m}^{(g)}, \theta) Pr(\boldsymbol{b_m}^{(g)} | \boldsymbol{\theta})}{\sum_{g=1}^{G} Pr(T_m \geq k | \boldsymbol{b_m}^{(g)}, \boldsymbol{\theta}) Pr(\boldsymbol{Y}_m | \boldsymbol{b_m}^{(g)}, \theta) Pr(\boldsymbol{b_m}^{(g)} | \boldsymbol{\theta})}
\end{aligned}
\tag{3.7}
$$

where $\boldsymbol{\theta}$ is the simulation parameters, and $\boldsymbol{Y}_m$ is the biomarker observations of subject m. Here $\boldsymbol{b_m}$ follows a mixture normal distribution with J components, and J is the number of latent classes. Since there is no closed form for calculating the integral in Equation 3.7, we simulate G sets of random effects from their distribution $Pr(\boldsymbol{b_m}|\boldsymbol{\theta})$, and average their effects to approximate this integral.

We apply the same simulation parameters as in setting (2), Table 3.5. The BSs and PARs are stored in Table 3.8 and Table 3.9 respectively. Similar to Table 3.3 and Table 3.6, the mean losses of BS in setting (3) are not significantly away from 0. However, the mean losses of BS in setting (3) are smaller than mean losses of corresponding BS in setting (1) and (2). This is because when data are simulated from setting (3), biomarker trajectories and times-to-aGVHD share both the latent class and random

Table 3.8: Brier Score of the pattern mixture model under simulation setting (3)

Prediction at week	1	2	3	4	5	6	7	8	9	10
Scenario 1										
$E(BS^{JM})$	0.317	0.433	0.494	0.505	0.534	0.549	0.548	0.586	0.607	0.713
$E(BS^{PM})$	0.320	0.423	0.472	0.476	0.501	0.513	0.509	0.549	0.568	0.676
$E(BS^{JM} - BS^{PM})$	-0.003	0.011	0.022	0.030	0.034	0.036	0.039	0.038	0.039	0.038
$SD(BS^{JM} - BS^{PM})$	0.018	0.019	0.020	0.022	0.025	0.028	0.033	0.037	0.044	0.053
Scenario 2										
$E(BS^{JM})$	0.314	0.429	0.493	0.503	0.531	0.546	0.545	0.582	0.601	0.713
$E(BS^{PM})$	0.318	0.417	0.468	0.472	0.495	0.508	0.504	0.543	0.561	0.675
$E(BS^{JM} - BS^{PM})$	-0.004	0.012	0.025	0.032	0.036	0.038	0.041	0.039	0.040	0.038
$SD(BS^{JM} - BS^{PM})$	0.019	0.020	0.021	0.024	0.026	0.029	0.034	0.038	0.045	0.055
Scenario 3										
$E(BS^{JM})$	0.327	0.436	0.491	0.502	0.529	0.545	0.545	0.580	0.600	0.709
$E(BS^{PM})$	0.327	0.424	0.469	0.473	0.496	0.508	0.507	0.542	0.563	0.673
$E(BS^{JM} - BS^{PM})$	0.000	0.012	0.022	0.029	0.033	0.036	0.038	0.038	0.038	0.035
$SD(BS^{JM} - BS^{PM})$	0.018	0.020	0.022	0.024	0.025	0.029	0.033	0.037	0.043	0.049
Scenario 4										
$E(BS^{JM})$	0.329	0.433	0.498	0.492	0.529	0.541	0.526	0.581	0.568	0.691
$E(BS^{PM})$	0.329	0.420	0.475	0.461	0.494	0.504	0.483	0.540	0.523	0.644
$E(BS^{JM} - BS^{PM})$	0.000	0.013	0.023	0.031	0.035	0.037	0.043	0.040	0.045	0.046
$SD(BS^{JM} - BS^{PM})$	0.017	0.019	0.019	0.022	0.025	0.028	0.034	0.038	0.046	0.058
Scenario 5										
$E(BS^{JM})$	0.330	0.428	0.497	0.491	0.528	0.541	0.525	0.578	0.568	0.691
$E(BS^{PM})$	0.329	0.415	0.469	0.456	0.488	0.498	0.477	0.532	0.521	0.645
$E(BS^{JM} - BS^{PM})$	0.002	0.013	0.028	0.036	0.040	0.043	0.047	0.046	0.047	0.046
$SD(BS^{JM} - BS^{PM})$	0.018	0.019	0.019	0.022	0.025	0.029	0.035	0.038	0.046	0.060
Scenario 6										
$E(BS^{JM})$	0.337	0.439	0.499	0.491	0.528	0.540	0.525	0.579	0.569	0.690
$E(BS^{PM})$	0.335	0.425	0.475	0.460	0.493	0.502	0.481	0.536	0.526	0.647
$E(BS^{JM} - BS^{PM})$	0.002	0.015	0.023	0.031	0.035	0.038	0.044	0.043	0.044	0.043
$SD(BS^{JM} - BS^{PM})$	0.018	0.019	0.020	0.023	0.025	0.030	0.034	0.038	0.044	0.057
Scenario 7										
$E(BS^{JM})$	0.256	0.382	0.467	0.473	0.518	0.535	0.521	0.578	0.569	0.692
$E(BS^{PM})$	0.246	0.366	0.445	0.442	0.484	0.499	0.478	0.537	0.522	0.644
$E(BS^{JM} - BS^{PM})$	0.010	0.016	0.022	0.031	0.034	0.037	0.043	0.041	0.047	0.048
$SD(BS^{JM} - BS^{PM})$	0.013	0.017	0.018	0.021	0.023	0.027	0.033	0.037	0.048	0.060

effects, leading to a stronger connection between the repeated biomarkers and times-to-aGVHD. The stronger the connection is, the relatively better the pattern mixture model performs. When comparing among various scenarios, we find the same pattern of BSs as in previous two settings.

The mean losses of PAR in setting (3) are not significantly away from 0, and they are smaller than mean losses of corresponding PAR in setting (1) and (2), same as BS. When comparing among various scenarios, we found the same pattern of PARs

Table 3.9: PAR of the pattern mixture model under simulation setting (3)

Predictionatweek	1	2	3	4	5	6	7	8	9	10
Scenario 1										
$E(PAR^{JM})$	0.417	0.551	0.610	0.623	0.647	0.661	0.659	0.690	0.704	0.786
$E(PAR^{PM})$	0.426	0.541	0.591	0.595	0.617	0.628	0.625	0.658	0.671	0.760
$E(PAR^{JM} - PAR^{PM})$	-0.009	0.010	0.019	0.028	0.030	0.033	0.034	0.033	0.033	0.027
$SD(PAR^{JM} - PAR^{PM})$	0.037	0.035	0.032	0.036	0.038	0.042	0.049	0.052	0.062	0.068
Scenario 2										
$E(PAR^{JM})$	0.417	0.548	0.610	0.620	0.643	0.656	0.655	0.684	0.697	0.785
$E(PAR^{PM})$	0.427	0.539	0.589	0.593	0.614	0.624	0.621	0.654	0.666	0.762
$E(PAR^{JM} - PAR^{PM})$	-0.010	0.009	0.021	0.027	0.030	0.032	0.034	0.031	0.031	0.023
$SD(PAR^{JM} - PAR^{PM})$	0.039	0.033	0.032	0.037	0.040	0.044	0.050	0.053	0.065	0.070
Scenario 3										
$E(PAR^{JM})$	0.430	0.553	0.606	0.618	0.641	0.654	0.654	0.683	0.698	0.785
$E(PAR^{PM})$	0.436	0.543	0.587	0.593	0.614	0.625	0.624	0.654	0.668	0.759
$E(PAR^{JM} - PAR^{PM})$	-0.006	0.010	0.019	0.025	0.027	0.029	0.030	0.028	0.030	0.026
$SD(PAR^{JM} - PAR^{PM})$	0.037	0.034	0.034	0.036	0.037	0.042	0.049	0.052	0.064	0.065
Scenario 4										
$E(PAR^{JM})$	0.439	0.552	0.614	0.611	0.642	0.653	0.640	0.686	0.672	0.768
$E(PAR^{PM})$	0.441	0.541	0.594	0.583	0.610	0.620	0.601	0.652	0.633	0.732
$E(PAR^{JM} - PAR^{PM})$	-0.001	0.011	0.019	0.028	0.033	0.033	0.039	0.034	0.039	0.036
$SD(PAR^{JM} - PAR^{PM})$	0.037	0.034	0.032	0.035	0.040	0.045	0.052	0.051	0.066	0.078
Scenario 5										
$E(PAR^{JM})$	0.442	0.547	0.613	0.609	0.640	0.652	0.638	0.682	0.672	0.769
$E(PAR^{PM})$	0.444	0.540	0.594	0.582	0.609	0.619	0.599	0.648	0.630	0.734
$E(PAR^{JM} - PAR^{PM})$	-0.002	0.007	0.019	0.027	0.031	0.032	0.040	0.034	0.042	0.035
$SD(PAR^{JM} - PAR^{PM})$	0.038	0.033	0.031	0.036	0.039	0.043	0.050	0.052	0.067	0.077
Scenario 6										
$E(PAR^{JM})$	0.450	0.556	0.614	0.609	0.639	0.650	0.636	0.682	0.671	0.765
$E(PAR^{PM})$	0.450	0.547	0.596	0.585	0.611	0.620	0.600	0.649	0.634	0.733
$E(PAR^{JM} - PAR^{PM})$	0.000	0.010	0.018	0.024	0.028	0.030	0.036	0.033	0.037	0.032
$SD(PAR^{JM} - PAR^{PM})$	0.036	0.033	0.032	0.035	0.038	0.043	0.052	0.052	0.066	0.074
Scenario 7										
$E(PAR^{JM})$	0.358	0.496	0.583	0.590	0.631	0.648	0.636	0.684	0.674	0.771
$E(PAR^{PM})$	0.339	0.477	0.562	0.564	0.601	0.615	0.598	0.651	0.634	0.735
$E(PAR^{JM} - PAR^{PM})$	0.019	0.019	0.021	0.027	0.030	0.033	0.038	0.033	0.041	0.037
$SD(PAR^{JM} - PAR^{PM})$	0.034	0.035	0.032	0.036	0.038	0.044	0.050	0.055	0.070	0.080

as in previous two settings.

3.5 Discussion

In this chapter, we introduce the pattern mixture model to make dynamic predictions of time-to-aGVDH with repeated biomarkers. The pattern mixture model identifies the patterns, or in other words, the latent classes of aGVHD, based on solely the times-to-aGVHD, and then summarizes the features of repeated biomark-

ers within each pattern. Our simulation study has shown that, when the patients' repeated biomarkers are strongly correlated with their times-to-aGVHD, pattern mixture model perform reasonably good.

We examine how the pattern mixture model performs when patients are of various risk classes of aGVHD. We consider two settings: there are inherent eleven latent classes, or four latent classes, as discussed in Chapter II. When patients are from eleven latent classes, their times-to-aGVHD are highly correlated with the repeated biomarker process. Under this setting, the pattern mixture model introduces a small loss to the accuracy of the time-to-aGVHD prediction, and approximately 3 out of 100 patients will be miss-predicted of their time-to-aGVHD. When patients are from four latent classes of aGVHD, the BS and PAR of the pattern mixture model decrease. However, it shows a stable loss of prediction accuracy as in setting (1).

In a more sophisticated setting, we assume the time-to-aGVHD not only depends on the latent class membership, but also on the individual deviations of biomarker trajectories. In this case, the pattern mixture model performs consistently good, even better than that in setting (2). This is because the shared random effects reinforce the correlation between biomarkers and the time-to-aGVHD, making the time-to-aGVHD itself as a stronger indicator of aGVHD latent class.

We also find that the overall prediction is better when the biomarker trajectories of patients between groups are more distinguishable with small variance of random effects. Moreover, larger measurement errors worsen prediction when available biomarker information is insufficient, but this influence is diluted when more biomarkers are collected. We also observe that when the random effects are negative correlated, the prediction improves.

Therefore, we conclude that the pattern mixture model can be applied to the dataset including both longitudinal biomarkers and times-to-aGVHD. It introduces small loss of prediction accuracy but it is more flexible and convenient to execute.

CHAPTER IV

Generalized Pattern Mixture Model in the Prediction of Time-to-Event

4.1 Introduction

Biomarkers have been applied in medical practice to accelerate disease diagnosis, monitor patients' health conditions, and predict treatment effects (Naylor, 2003; Mayeux, 2004). Many recent studies have demonstrated that biomarker profiles differ by person, and this difference cannot be fully explained by individual-level random effects only. For example, Proust-Lima et al. (2016) found that there are four subgroups of dementia patients, and each group has a distinct pattern of how the semantic memory changes over time.

Each risk group might have different risks of adverse events, and thus the length of follow-up also differs by risk group. In this case, identifying the risk group is important; otherwise, the population from the low risk group will be over-represented given the fact that they have more biomarker observations. In Chapter III, we introduced how to apply the pattern mixture model to identify different biomarker trajectories in patients who have received BMT. Patients' risk group membership was determined by their times-to-aGVHD. Patients in the same risk group had an equal number of

biomarker observations, and we made the assumption that they shared the same biomarkers distribution.

In Chapter III, patients' times-to-aGVHD were grouped in weeks, and we assumed patients who had developed aGVHD in the same week after BMT shared similar biomarker patterns. This setting can be generalized to other diseases. In practice, patients' onset of non-fatal disease might not be accurately recorded, especially when the disease is self-reported. Another example is the follow-up of patients with chronic diseases, whether a patient's time-to-event is recorded promptly depends on the timing and frequency of follow-up after hospital discharge. In these cases, it is reasonable to group the patients' times-to-event into discrete intervals. In practice, these time intervals might or might not be of equal length. For example, in a study of the elderly at risk for dementia, the biomarkers (semantic memory scores) were measured at baseline, year 1, 3, 5, 8, 10, and so forth (Proust-Lima et al., 2016).

In Chapter III, we showed how to use the pattern mixture model to a setting, in which patients were either free of aGVHD or their onset of aGVHD was completely recorded. Patients receiving BMT are under close monitoring. Moreover, aGVHD is a complication typically happening within 100 days after BMT. In this case, we do not have missing observations in patients' times-to-aGVHD or biomarker observations. However, in other clinical studies, the event time might not be fully observed due to administrative censoring, loss to follow up, or competing risks. For chronic diseases, it might take years to observe the progression of a disease or death, so that the missingness of event times is quite common. The censored participants still contribute to the analysis by the partial information that they are event-free at the time of censoring. Classic methods, such as parametric survival regression and semi-parametric Cox regression, utilize this partial information and achieve larger

power compared with the complete-case method. Moreover, when the event time is not missing completely at random, complete cases are a biased sample of the original dataset. Thus, the complete case analysis generates biased parameter estimation.

There is another type of missingness in event times, when the patients are "cured." For example, Rama et al. (2010) found that if patients with locally advanced breast cancer survived seven years after their cancer diagnosis, they were considered to be cured and would be less likely to die from breast cancer. Dal Maso et al. (2014) found that among patients with thyroid, testis or corpus uteri cancer, the cured fraction was as high as 90%. These cured patients might share different clinical characteristics than susceptible patients; thus, it is important to identify the cured patients and model their biomarker patterns separately.

In this chapter, we will discuss how to generalize this pattern mixture model in the prediction of time-to-event, considering both the cured fraction of the population and random censoring in the event time. The key improvement in this project is that we utilize the information from the censored patients, estimate their probability of being cured, and refine the parameter estimation in the longitudinal biomarker process for both cured and susceptible patients.

The rest of this chapter is organized as follows. First, we will introduce how to structure the model and make predictions with the generalized pattern mixture model (GPMM). Second, we simulate data with independent and dependent censoring, and evaluate the performance of this GPMM compared to the complete-case pattern mixture model. We conclude the chapter with a discussion.

4.2 Methods

Similar to Chapter III, we define T_i to be the recorded last follow-up in weeks for subject i, which is the minimum of time-to-event in weeks for subject i, T_i^*, or the censoring time S_i. We also define δ_i as the indicator of whether subject i has experienced aGVHD ($\delta_i = 1$) or has been censored ($\delta_i = 0$).

We define h_i as the group indicator of participant i, and $h_i = T_i^*$ if the patient develops aGVHD T_i^* weeks after BMT. Given there are fewer patients who develop aGVHD after 80 days, we combine all the aGVHD cases ten weeks after BMT and define that they all belong to group 10. The "cured" patients, who will never develop aGVHD after BMT, will join group 11. Unlike the aGVHD data in Chapter III, h_i is partially observed because T_i^* is only partially observed. We cannot tell whether an early censored patient is aGVHD-free or will develop aGVHD later.

We also define $\boldsymbol{Y_i} = (Y_i(t_1), Y_i(t_2), \cdots, Y_i(t_{n_i}))$ as the biomarker history of subject i at time $(t_1, t_2, \cdots, t_{n_i})$, where n_i is the total number of biomarker observations for subject i, for $i = 1, 2, \cdots, n$. For simplicity, we assume that the continuous biomarker observations $\boldsymbol{Y_i}$ change linearly over time, but it is straightforward to generalize them as other functional forms of time. Random effects $\boldsymbol{b_i} = (b_{0i}, b_{1i})$ are introduced to reflect the individual deviation of the biomarker trajectory from the population mean, where $\boldsymbol{b_i}$ follows a Gaussian distribution with mean $\mathbf{0}$ and variance-covariance D.

We assume the mean biomarker trajectory depends on the patient's time-to-aGVHD. The susceptible patients are naturally grouped by their times-to-aGVHD week until day 100 after BMT. Patients experiencing aGVHD within the same week share the same biomarker pattern, therefore aGVHD-free patients in group 11 will share a distinct biomarker pattern than those susceptible patients.

Let $B_i = (\boldsymbol{Y_i}, T_i, \delta_i)$ represent the observed data for subject i, and $\boldsymbol{O_i} = (\boldsymbol{Y_i}, h_i)$ denote the complete data, in which h_i are not fully observed.

We assume the biomarker observations, $\boldsymbol{Y_i}$, follow a Gaussian distribution such that:

$$\boldsymbol{Y_i}|h_i = h = \boldsymbol{X_i}\boldsymbol{\beta^{(h)}} + \boldsymbol{Z_i}\boldsymbol{b_i} + \boldsymbol{\epsilon_i} \tag{4.1}$$

$$Y_i(t_j)|h_i = h = \beta_0^{(h)} + \beta_1^{(h)}t_j + b_{0i} + b_{1i}t_j + \epsilon_i(t_j) \tag{4.2}$$

where $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ is the design matrix for subject i, and the measurement error $\boldsymbol{\epsilon_i}$ follows a Gaussian distribution with mean $\mathbf{0}$ and variance $\sigma^2\boldsymbol{I_{n_i}}$. Patients in each risk group h share the same mean profile of biomarker, denoting by group-specific parameters $\beta^{(h)}$.

Define p_h as the marginal probability of a patient coming from group h, specifically, the patient has experienced aGVHD after the hth week, or this patient is aGVHD-free if $h = 11$.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \cdots, \boldsymbol{\beta}^{(11)}), D, \sigma^2)$ denote the parameters involved in the longitudinal biomarker process, and $\boldsymbol{\xi} = (\boldsymbol{\theta}, \boldsymbol{\pi} = (p_1, p_2, \cdots, p_{11}))$ be the complete parameter set. The likelihood function given the complete data $\boldsymbol{O} = (\boldsymbol{Y}_i, h_i)$ is:

$$\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{O}) = \prod_{i=1}^{n}\prod_{h=1}^{11}[p_h f(\boldsymbol{Y_i}|\boldsymbol{\beta}^{(h)}, D, \sigma^2)]^{I(h_i=h)}$$

The corresponding log-likelihood function is:

$$\begin{aligned} l(\boldsymbol{\xi}|\boldsymbol{O}) &= \sum_{i=1}^{n}\sum_{h=1}^{11}[I(h_i = h)log(p_h) + I(h_i = h)log(f(\boldsymbol{Y_i}|\boldsymbol{\beta}^{(h)}, D, \sigma^2))] \\ &= l_1(p_h|\boldsymbol{O}) + l_2(\boldsymbol{\theta}|\boldsymbol{O}) \end{aligned} \tag{4.3}$$

where $l_1(p_h|\boldsymbol{O}) = \sum_{i=1}^{n}\sum_{h=1}^{11} I(h_i = h)log(p_h)$ involving only the marginal probability of latent classes, and $l_2(\boldsymbol{\theta}|\boldsymbol{O}) = \sum_{i=1}^{n}\sum_{h=1}^{11} I(h_i = h)log(f(\boldsymbol{Y_i}|\boldsymbol{\beta}^{(h)}, D, \sigma^2))$ involving only the biomarker distribution. Thus, the log-likelihood function in Equation 4.3 is separated into two parts involving parameter p_h and $\boldsymbol{\theta}$ respectively. Given the complete data $\boldsymbol{O}$, it would be straightforward to achieve the MLE of $\boldsymbol{\xi}$; however, with only the observed data $\boldsymbol{B}$, we cannot construct the likelihood function in a closed form. One solution is using the Expectation-Maximization (EM) algorithm to facilitate the parameter estimation process.

4.2.1 Model fitting

The EM algorithm starts with computing the expectation of unobserved data, and maximize the likelihood function given these expectations. In our setting, patients' true time-to-aGVHD T_i^*, or equivalent group indicator h_i, are only partially observed, so we will begin with obtaining the expectation of unobserved h_i.

For an observed cured patient, the group indicator h_i is 11; and $h_i = T_i^*$ for a patient who had experienced an event. For a censored patient, we need to calculate the expected group membership h_i. Given the log-likelihood in Equation 4.3, at the

$q+1$ step, the expectation of h_i is:

$$E(h_i) = \begin{cases} 11, & T_i^* = \infty \\ T_i^*, & \delta_i = 1 \\ \dfrac{p_h^{(q)} f(\boldsymbol{Y_i}|\boldsymbol{\beta}^{(h)(q)}, D^{(q)}, \sigma^{2(q)})}{\sum_{h=c}^{11} p_h^{(q)} f(\boldsymbol{Y_i}|\boldsymbol{\beta}^{(h)(q)}, D^{(q)}, \sigma^{2(q)})}, & \delta_i = 0, T_i = c \end{cases} \tag{4.4}$$

For censored patients, the expectation of group indicator h_i equals the weighted probability of being in group h given the patient's follow-up time and biomarker observations. Thus, the estimated probability of patient i's group membership, $\boldsymbol{p_i} = (p_{1i}, p_{2i}, \cdots, p_{11,i})$, could be calculated as:

$$\boldsymbol{p_i} = \begin{cases} (0, \cdots, 1), & T_i^* = \infty \\ (0, \cdots, 0, p_{ci} = 1, 0, \cdots 0,), & \delta_i = 1, T_i = c \\ (0, \cdots, 0, p_{ci} = \dfrac{p_h^{(q)} f(\boldsymbol{Y_i}|\boldsymbol{\beta}^{(h)(q)}, D^{(q)}, \sigma^{2(q)})}{\sum_{h=c}^{11} p_h^{(q)} f(\boldsymbol{Y_i}|\boldsymbol{\beta}^{(h)(q)}, D^{(q)}, \sigma^{2(q)})}, \cdots), & \delta_i = 0, T_i = c \end{cases}$$

To facilitate the parameter estimation in the biomarker process, we obtain the expectation of unobserved random effects $\boldsymbol{b_i}$ and measure errors $\boldsymbol{\epsilon_i}$ as in Chapter 2.2.2. Define $\boldsymbol{H}^{(q)} = Z_i D^{(q)} Z_i' + \sigma^{2(q)} \boldsymbol{I_{n_i}}$, and given the joint distribution of $(\boldsymbol{Y_i}, \boldsymbol{b_i}, \boldsymbol{\epsilon_i})$

we obtain:

$$
\begin{aligned}
E(\boldsymbol{b_i b_i'}|\boldsymbol{Y_i}, \boldsymbol{\xi^{(q)}}) &= E_{\boldsymbol{b_i}|\boldsymbol{Y_i}}(\boldsymbol{b_i}|\boldsymbol{Y_i}) E_{\boldsymbol{b_i}|\boldsymbol{Y_i}}(\boldsymbol{b_i}|\boldsymbol{Y_i})' + cov_{\boldsymbol{b_i}|\boldsymbol{Y_i}}(\boldsymbol{b_i}|\boldsymbol{Y_i}) \\
E(\boldsymbol{e_i e_i'}|\boldsymbol{Y_i}, \boldsymbol{\xi^{(q)}}) &= E_{\boldsymbol{e_i}|\boldsymbol{Y_i}}(\boldsymbol{e_i}|\boldsymbol{Y_i}) E_{\boldsymbol{e_i}|\boldsymbol{Y_i}}(\boldsymbol{e_i}|\boldsymbol{Y_i})' + tr\{cov_{\boldsymbol{e_i}|\boldsymbol{Y_i}}(\boldsymbol{e_i}|\boldsymbol{Y_i})\} \\
\text{where} \quad E_{\boldsymbol{b_i}|\boldsymbol{Y_i}}(\boldsymbol{b_i}|\boldsymbol{Y_i}, h_i = h) &= D^{(q)} Z_i \boldsymbol{H^{(q)}}^{-1} (\boldsymbol{Y_i} - X_i \boldsymbol{\beta}^{(h)}) \\
cov_{\boldsymbol{b_i}|\boldsymbol{Y_i}}(\boldsymbol{b_i}|\boldsymbol{Y_i}, h_i = h) &= D^{(q)} - D^{(q)} Z_i \boldsymbol{H^{(q)}}^{-1} Z_i D^{(s)} \\
E_{\boldsymbol{e_i}|\boldsymbol{Y_i}}(\boldsymbol{e_i}|\boldsymbol{Y_i}, h_i = h) &= \sigma^{2(q)} \boldsymbol{H^{(q)}}^{-1} (\boldsymbol{Y_i} - X_i \boldsymbol{\beta}^{(h)}) \\
cov_{\boldsymbol{e_i}|\boldsymbol{Y_i}}(\boldsymbol{b_i}|\boldsymbol{Y_i}, h_i = h) &= \sigma^{2(q)} (\boldsymbol{I_{n_i}} - \sigma^2 \boldsymbol{H^{(q)}}^{-1})
\end{aligned}
$$

With the complete data sufficient statistics $(\boldsymbol{h_i}, \boldsymbol{b_i b_i'}, \boldsymbol{e_i e_i'})$, we can compute the expectation of log-likelihood $l_1(p_h|\boldsymbol{O})$ and $l_2(\boldsymbol{\theta}|\boldsymbol{O})$. In the $q+1$ iteration and the M-step, we can achieve the MLE for parameters $\boldsymbol{\xi}^{(q+1)} = (p_h^{(q+1)}, \boldsymbol{\theta}^{(q+1)})$ by maximizing the corresponding expectation of log-likelihood. The MLE of p_h is:

$$
\widehat{p_h} = \frac{1}{n} \sum_{i=1}^{n} p_{hi} \tag{4.5}
$$

The MLEs of parameters in longitudinal biomarker process, $\boldsymbol{\theta}^{(q+1)}$, are:

$$
\begin{aligned}
\widehat{D} &= \frac{1}{n} \sum_{i=1}^{n} \Big[\sum_{h=1}^{11} p_{hi} E(\boldsymbol{b_i}|\boldsymbol{Y_i}, h_i = h) E(\boldsymbol{b_i}|\boldsymbol{Y_i}, h_i = h)' + cov_{\boldsymbol{b_i}|\boldsymbol{Y_i}}(\boldsymbol{b_i}|\boldsymbol{Y_i}) \Big] \\
\widehat{\sigma^2} &= \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} \Big[\sum_{h=1}^{11} p_{hi} E(\boldsymbol{e_i}|\boldsymbol{Y_i}, h_i = h) E(\boldsymbol{e_i}|\boldsymbol{Y_i}, h_i = h)' + tr\{cov_{\boldsymbol{e_i}|\boldsymbol{Y_i}}(\boldsymbol{e_i}|\boldsymbol{Y_i})\} \Big] \\
\widehat{\boldsymbol{\beta}}^{(h)} &= \Big[\sum_{i=1}^{n} p_{hi} X_i' \widehat{\boldsymbol{H_i}^{-1}} X_i \Big]^{-1} \Big[\sum_{i=1}^{n} p_{hi} X_i' \widehat{\boldsymbol{H_i}^{-1}} Y_i \Big]
\end{aligned} \tag{4.6}
$$

We repeat the above E-step and M-step until we reach some pre-specified convergence criteria. Finally, we obtain the parameter estimations for the marginal cure probability p_h and the biomarker trajectories $\boldsymbol{\theta}$. With these parameter estimations, we could predict one patient's risk of events given his/her current available biomarker

records.

4.2.2 Prediction with Generalized Pattern Mixture Model

The marginal distribution of observed biomarkers is a finite mixture of Gaussian distribution, and the posterior probability of a new patient m developing aGVHD at week j, with the observed biomarkers history by week k $\boldsymbol{Y_m} = (Y_{m1}, Y_{m2}, \cdots, Y_{m,k})$, is that:

$$Pr(T_m = j|\boldsymbol{Y_m}, \boldsymbol{\xi}) \quad = \quad \frac{f(\boldsymbol{Y_m}|\boldsymbol{\beta}^{(j)}, D, \sigma^2)Pr(h_m = j)}{\sum_{l=k}^{11} f(\boldsymbol{Y_m}|\boldsymbol{\beta}^{(l)}, D, \sigma^2)P(h_m = l)} \tag{4.7}$$

for $j \geq k$. Bayes' theorem is applied here to compute the posterior probability of subject m experiencing the event in week j given his/her biomarker history $\boldsymbol{Y_m}$.

4.3 Simulation and Result

In this section, we evaluate the benefit of adjusting for the cured fraction in the pattern mixture model, compared with using only complete cases. We consider the scenarios with independent and dependent censoring. In practice, independent censoring is mainly caused by administrative censoring or random drop-out. Multiple reasons contribute to dependent censoring, such as competing risks or other adverse events. We compare the prediction performance between the generalized pattern mixture model and pattern mixture model with complete cases only in these two scenarios.

Data are simulated from a pattern mixture distribution with 11 patterns, representing time-to-event after week 1, week 2, $\cdots$, week 10, and the "cured" patients. Patients in each pattern share a distinct mean biomarker profile with individual-specified random effects. For the sake of simplicity, we assume the mean biomarker trajectory is a linear function of time. The independent censoring time is simulated

from a uniform (1.5,17), resulting in around 34% censoring. In the dependent censoring scenario, the censoring time is simulated from a uniform distribution range from 1.5 to 15 plus the true event time. So the probability of censoring depends on the true event time. This results in around 35% censoring.

In each simulation, a sample of 220 patients is generated as the training set, with their follow-up of biomarkers truncated by their longest follow-up time, the true time-to-event or censoring, whichever comes first. Another dataset of 220 patients is also generated from the same distribution and used as a test set. One thousand simulations are executed to evaluate the performance of the generalized pattern mixture model against complete-case pattern mixture model.

Similar to Chapter III, we use a dynamic Brier Score to summarize the performance of prediction. As a reminder, we define the dynamic BS as:

$$BS(k) = 1 - \frac{1}{\sum_{i=1}^{n} I(T_i \geq k)} \sum_{i=1}^{n} \sum_{j=1}^{J} I(T_i \geq k)[f_{ij}(\boldsymbol{Y_i}) - o_{ij}]^2 \tag{4.8}$$

where $o_{ij} = 1$ if patient i develops the event after week j, and $f_{ij}(\boldsymbol{Y_i})$ is the posterior probability of patient i developing the event after week j. Therefore, the BS is the sum of squared prediction errors across all subjects who are still at risk at prediction time k.

In the independent censoring scenario, we check the prediction accuracy of marginal probabilities of patterns, together with the dynamic BS changing over prediction times. First we evaluate the estimation of the marginal probability of each pattern. The marginal probability of patterns has certain medical implications. First, it reflects the structure of the targeted population. Second, it represents the risk of events

when biomarker observations are not available. The squared error of marginal probability estimation is measured by the Euclidean distance between the true marginal probability and the estimated one for each of the two methods. The lower the squared error is, the better the estimation of marginal probabilities.

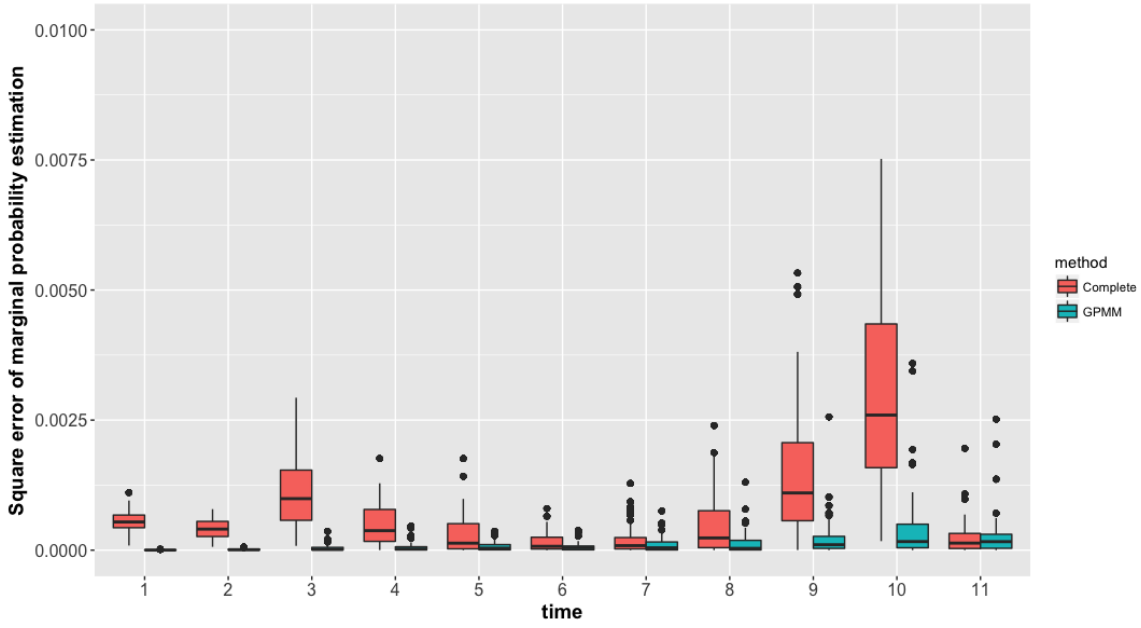

Figure 4.1: Squared error of marginal probability estimation of complete-case analysis (red) and GPMM (green)

As shown in Figure 4.1, GPMM controls the squared error well, while complete-case analysis loses efficiency in marginal probability estimation. This is because in this independent censoring scenario, each pattern has a different probability of censoring. The longer the true event time is, the more likely it is censored. Thus, the relative frequency of patterns in the complete cases does not reflect the true distribution of patterns in the target population. GPMM, in contrast, redistributes censored patients into different patterns. Therefore, it alleviates the effects of censoring in estimating the marginal probability.

Next, we examine the performance of prediction between the two models. Based on the results shown in Figure 4.2, the Brier Scores of the two methods is similar. GPMM achieves a higher BS in early predictions, but the effect size of this improve-

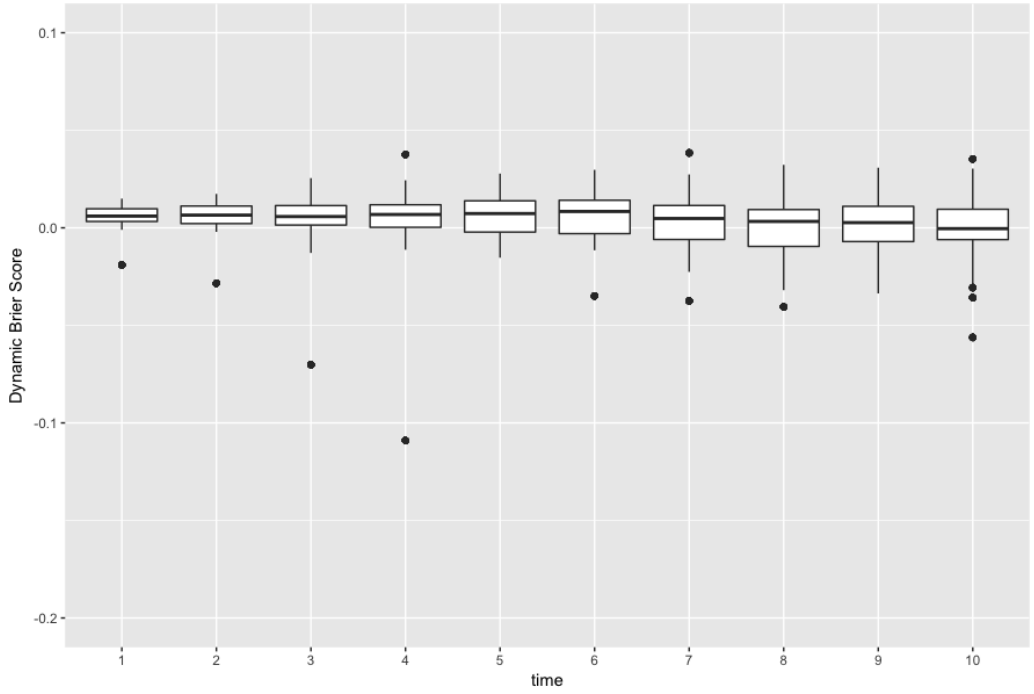

Figure 4.2: Differences of the dynamic Brier Score between GPMM and complete-case analysis

ment is small. With more biomarkers available, the difference of BS between the two methods decreases.

Table 4.1 lists the mean BS of GPMM, complete-case analysis, and the mean and the standard deviation of the difference between the two methods. We found that the complete-case analysis surpasses GPMM in later predictions when more biomarker observations are available. This is because complete-case analysis, although it has a reduced power due to the smaller sample size, provides unbiased estimation of biomarker profiles over time. In contrast, GPMM incorporates censored cases into each pattern with an adjusted weight, and results in a biased estimation of biomarker profiles. When there are fewer than 8 biomarker observations, the effect of biased

biomarker profile estimation is alleviated by more accurate marginal probabilities of patterns; thus, GPMM does better in early predictions. However, when there are

Table 4.1: Brier Scores of GPMM and complete-case analysis in the independent censoring scenario

Prediction at week	1	2	3	4	5	6	7	8	9	10
GPMM	0.146	0.247	0.375	0.484	0.587	0.675	0.729	0.773	0.828	0.873
Complete-case	0.140	0.241	0.369	0.479	0.584	0.674	0.732	0.780	0.835	0.882
Difference (mean)	0.006	0.006	0.005	0.005	0.002	0.001	-0.003	-0.007	-0.007	-0.009
Difference (SD)	0.005	0.008	0.013	0.019	0.031	0.039	0.042	0.055	0.061	0.064

more biomarker observations and fewer candidate patterns to predict, the effect of unbiased biomarker profile estimation dominates the biased estimation of marginal probabilities of patterns; thus, complete-case analysis does better in later predictions. Therefore, there is a trade-off effect of biased biomarker estimation and more accurate marginal probabilities of patterns.

In the dependent censoring scenario, the censoring time depends on the true event time. We also check the prediction accuracy of marginal probabilities of patterns and the dynamic BS changing over prediction time. As shown in Figure 4.3, GPMM controls the squared error well, while complete-case analysis has an obvious larger error in marginal probability estimation. This is because in the dependent censoring scenario, times-to-censoring depend on true times-to-event, so that complete cases are not a good representative of the target population. GPMM, however, estimates the pattern of censored patients and thus adjusts the marginal probability of patterns. Compared with the estimation error shown in Figure 4.1 when data are simulated under independent censoring, the estimation errors of marginal probability for both complete-case analysis and GPMM inflate.

Next, we check the performance of prediction between the two models. Based on the results shown in Figure 4.4, GPMM, on average, obtains a higher BS than the

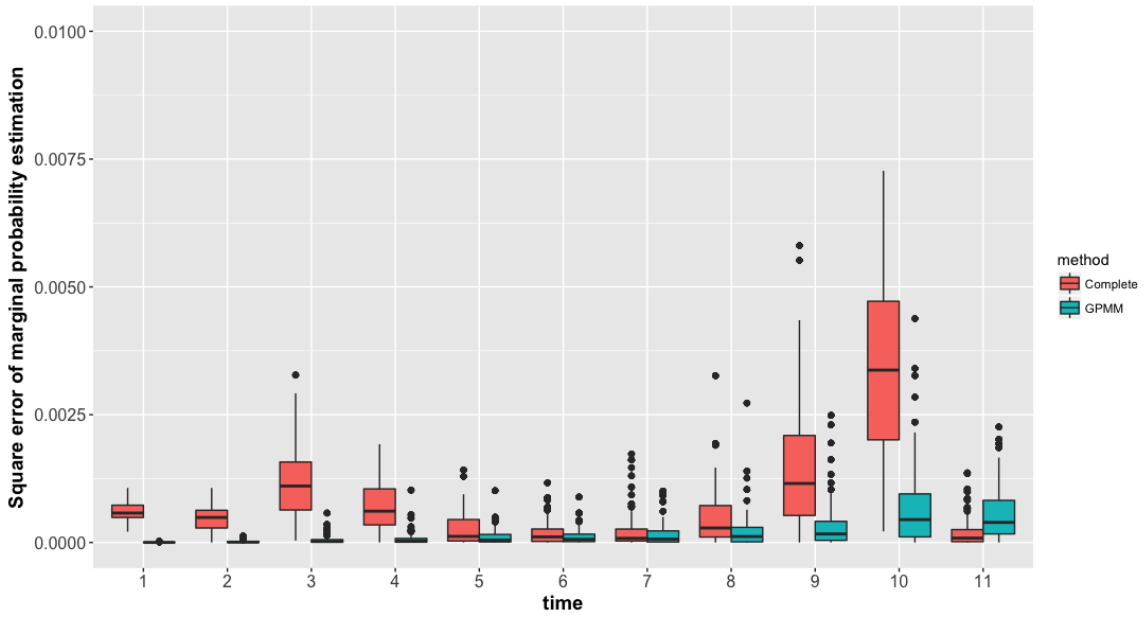

Figure 4.3: Squared error of marginal probability estimation of complete-case analysis (red) and GPMM (green)

complete-case analysis, no matter how many biomarker observations are available. However, the increase in BS is close to 0.

Table 4.2 presents the mean BS of GPMM, complete-case analysis, the mean and the standard deviation of the difference between the two methods. We found that

Table 4.2: Brier Scores of GPMM and complete-case analysis in the dependent censoring scenario

Prediction at week	1	2	3	4	5	6	7	8	9	10
GPMM	0.146	0.237	0.334	0.416	0.473	0.520	0.557	0.600	0.656	0.749
Complete-case	0.140	0.232	0.330	0.414	0.471	0.519	0.556	0.598	0.655	0.748
Difference (mean)	0.006	0.005	0.004	0.003	0.002	0.001	0.001	0.003	0.001	0.001
Difference (SD)	0.005	0.007	0.009	0.015	0.019	0.023	0.027	0.030	0.032	0.029

GPMM does consistently better than landmark analysis, but the mean difference of BS is relatively small compared to its standard deviation. In the independent censoring scenario, we find that the unbiased estimation of biomarker profiles offsets the biased estimation of marginal probabilities. Thus, complete-case analysis

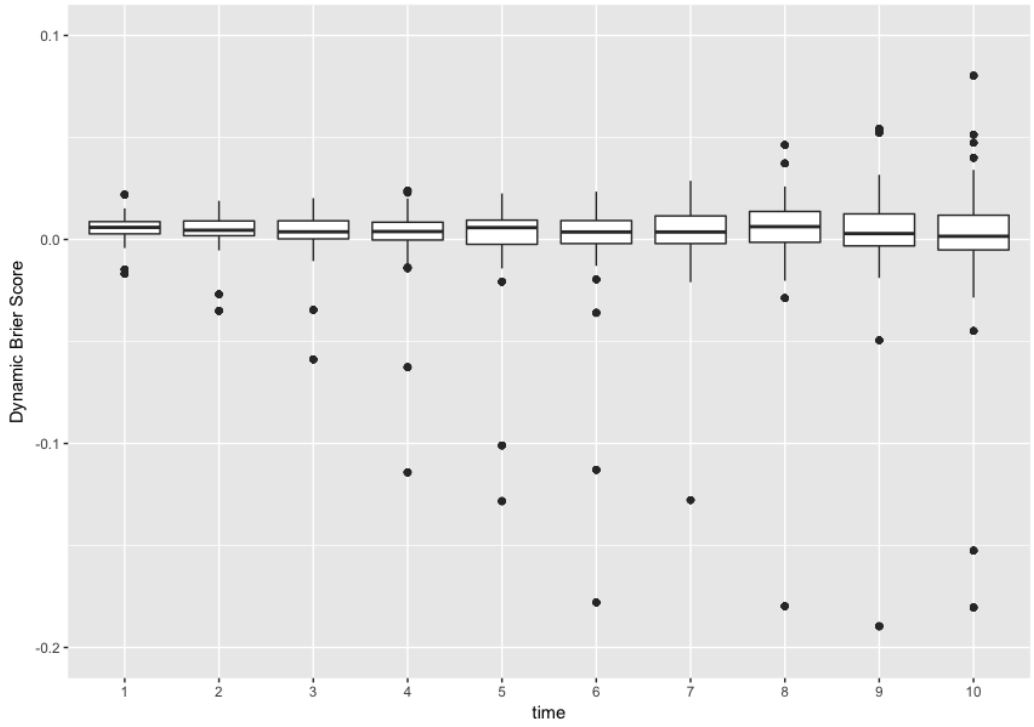

Figure 4.4: Differences of the dynamic Brier Score between GPMM and complete-case analysis

surpasses GPMM in later predictions. However, in the dependent censoring scenario, the marginal probability estimation in complete case analysis does not reflect the true distribution of patterns, and it produces more biased marginal probabilities than the true ones, as demonstrated by Figure 4.3. Thus, the complete-case analysis is more sensitive to censoring, especially dependent censoring. On the other hand, GPMM better predicts the marginal probabilities of patterns.

4.4 Discussion

In this Chapter, we introduce how to apply the pattern mixture model to the setting in which censored cases are allowed. In practice, censoring, or more generally speaking, partially observed information, is very common. Complete-case analysis, which ignores the censored cases, not only suffers from a lower power, but also results

in a biased marginal probability estimation. GPMM, however, reweights the censored cases and allows them to contribute to the model fitting given the information up to the censoring time. As a result, GPMM has a more accurate estimation of marginal probability. This benefit, as demonstrated in the simulation section, drives the improvement of GPMM over complete-cases analysis.

In this project, a sample of 220 patients are simulated as a training or test dataset. We use 220 rather than 200 in Chapter III because we want to maintain enough power for the complete-case analysis. With around 30% censoring, the complete-case analysis uses dataset of 150 patients from eleven patterns. On the other hand, GPMM has a much larger power due to its utilization of censored cases. Moreover, we require enough cases within each pattern to achieve a robust estimation of the mean pattern profile. However, in complete-case analysis with independent or dependent censoring, it is likely that there are only a limited number of cases in some patterns. Thus, if these cases are extreme cases, we might end up with biased estimations for biomarker profiles. As demonstrated in Figure 4.1 and Figure 4.3, there are more outliers of estimation in complete case than GPMM. Therefore, when there are only limited number of patients, or if some of patterns have few patients, complete-case analysis is not recommended.

CHAPTER V

Bootstrap Methods for Determining the Number of Latent Classes in Joint Modeling

5.1 Introduction

Joint modeling (JM) has been widely applied to medical studies in which both the longitudinal biomarkers and the time-to-event are of interest. One type of JM, joint modeling with latent classes (JMLC) has recently received extensive attention and has been applied to the prediction of prostate cancer (Lin et al., 2002), AIDS (Liu et al., 2015), severe hot flashes (Jiang et al., 2015) and dementia (Proust-Lima et al., 2016). JMLC assumes the population consists of individuals in various latent classes defined by the risk of the disease, and individuals from the same latent class share the same distribution of the longitudinal biomarker and the time-to-event. One prerequisite of applying JMLC is knowing the number of latent classes.

Although determining the number of latent classes in JMLC can be part of the empirical data analysis when there is enough evidence in the data to show that multiple classes exist, prior knowledge regarding the exact number of classes is usually unavailable. Generally speaking, we expect an adequate number of latent classes to capture the heterogeneity of the biomarker patterns and the time-to-event distribu-

tions in the population, while avoiding redundant components so that we can control the overall complexity of JMLC.

Akaike information criterion (AIC) and Bayesian information criterion (BIC) are current standard approaches used to choose the number of latent classes in JMLC (Proust-Lima et al., 2014; Liu et al., 2015; Rouanet et al., 2016). These information-based criteria are all measures of goodness of model fitting and add a penalty term for complexity to the negative log-likelihood. The penalty term, $2k$ for AIC, and $klog(n)$ for BIC, is a function of the effective number of parameters k and the sample size n, so that the final selected model achieves both a good fit to the data and parsimony (Leroux, 1992; Keribin, 2000; Akogul and Erisoglu, 2016). The model with the smallest value of AIC or BIC is preferred.

These information-based criteria quantify whether it is worth having a richer model in terms of information gain. However, it is hard to quantify how much practical information gain is associated with a one-unit change in BIC. Moreover, AIC has been criticized for prefering models that contain more latent classes than the actual number (Olofsen and Dahan, 2013). In general, the choice of the number of latent classes should not only be based on the smallest information criterion, but also on meaningful latent classes and a good discrimination between each latent class (Proust-Lima et al., 2014). Therefore, this study considers other methods for choosing the number of latent classes in JMLC.

JMLC is inspired by finite mixture modeling (Vermunt and Magidson, 2003; Proust-Lima et al., 2014), so it is beneficial to review the methods for selection of the number of latent classes in mixture models. Other than the information-based criteria, another accepted method is regularization, which adds one or more penalty

terms to the negative log-likelihood used for estimation. These penalty terms could be functions of the probabilities of latent classes, or functions that summarize the similarity between latent classes. The regularization method usually starts with an adequately large number of latent classes, and then reduces the number by merging latent classes with similar distributions of outcomes of interest together, and eliminates latent classes with too few subjects (Chen and Khalili, 2008; Lindsten et al., 2011). For example, consider a one-dimensional location mixture model with location parameters $\boldsymbol{\mu}$, where $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_J)$ are the location parameters for latent classes $1, 2, \cdots, J$. A penalty function of difference of location parameters $\lambda \sum_{i \neq j} f(\mu_i - \mu_j)$ is added to the log-likelihood, where λ is the tuning parameter that larger λ results in smaller number of latent classes. This penalty function forces similar location parameters to be equal, thus reducing the number of latent classes.

Another potential method for selection of the number of latent classes is through hypothesis testing, such as the score test or likelihood ratio test (LRT) (Neyman and Scott, 1965; Lindsay, 1995), which are extensively used in parametric model selection. Hypothesis testing starts with setting the null and alternative hypotheses for the number of latent classes, and then calculates the null test statistic based on the data. The p-value is calculated and used as the evidence to support or reject the null hypothesis. Compared with AIC, BIC and regularization, the hypothesis testing method is an inference process and quantifies the information in the data (p-value) of choosing the number of latent classes.

The development and comparison of methods for selecting the number of latent classes in mixture models is focused primarily on the mixtures of Gaussian distributions (Lo et al., 2001; Akogul and Erisoglu, 2016). For example, Lo et al. (2001) derived the asymptotic distribution of the LRT statistic testing the number of latent

classes when the sample is drawn from a mixture of normal distributions. Akogul and Erisoglu (2016) compared the performance of AIC and BIC to each other with mixtures of multivariate normal distributions. The majority of practical analyses use information-based criteria to select the number of latent classes in mixture models (Fonseca and Cardoso, 2007; Akogul and Erisoglu, 2016; Zhang, 2016; Mehrjou et al., 2016), because the information-based criteria are comparatively easy to compute. Tein et al. (2013) reviewed 38 published papers using latent profile analysis and found that BIC is the most accepted criterion in model selection.

In this project, we propose a hypothesis testing method to select the number of latent classes in JMLC. However, a LRT or score test should be applied with caution in the latent class setting, as one of the regularity conditions for standard asymptotics is violated. This condition requires that if the the log-likelihood function is maximized at the true parameter $\boldsymbol{\xi}_0$, then $\boldsymbol{\xi}_0$ should be an interior point in its support. Let us consider a test of $\mathcal{H}_0 : J = 3$ vs. $\mathcal{H}_1 : J = 4$, where J is the number of latent classes. Two models are fitted assuming the outcome $\boldsymbol{Y}$ comes from the corresponding distributions:

$$\begin{aligned} \boldsymbol{Y} &\sim \sum_{j=1}^{3} \pi_j f(.|\boldsymbol{\xi_j}) \\ \boldsymbol{Y} &\sim \sum_{j=1}^{4} \pi_j f(.|\boldsymbol{\xi_j}) \end{aligned}$$

where π_j is the marginal probability that $\boldsymbol{Y}$ is a member of latent class j, and $\boldsymbol{\xi_j} \in \boldsymbol{\Xi^p}$ is the p-vector of parameters in latent class j. Therefore the hypothesis of interest is equivalent to $\mathcal{H}_0 : \pi_4 = 0, \boldsymbol{\xi_4} \neq \boldsymbol{\xi_l}$ vs. $\mathcal{H}_1 : \pi_4 \neq 0, \boldsymbol{\xi_4} \neq \boldsymbol{\xi_l}$, for $l = 1, 2, 3$. The null value of π_4 lies on the boundary of its support $[0, 1]$, so the asymptotic distribution of the LRT statistic is not a χ^2 distribution with $p + 1$ degrees of freedom.

Some research has developed the theoretical asymptotic null distribution of the LRT statistic when the null hypothesis contains boundary values (Lo et al., 2001; Crainiceanu and Ruppert, 2004; Stram and Lee, 1994). Lo et al. (2001) proved that the LRT statistic, in a mixture of normal distributions testing k_0 components against the alternative normal mixture distribution with k_1 components, has a null distribution that is a weighted sum of χ^2_1 variables, which is not available in a closed form. Similar conclusions are drawn from studies on likelihood ratio testing for zero variance components in linear mixed models (Stram and Lee, 1994; Crainiceanu and Ruppert, 2004). Moreover, the joint distribution of the longitudinal biomarker observations and time-to-event has a more complicated form than a mixture of normal distributions or longitudinal data. Though using LRT to select the number of latent classes in JMLC seems plausible, this approach has been limited use due to the difficulty in deriving the theoretic asymptotic null distribution of the LRT statistic.

As an alternative to deriving the theoretic null distribution of the LRT statistic, we propose using the parametric bootstrap. Some research has discussed using the bootstrap with the LRT for determination of the number of latent classes (McCutcheon, 1987; McLachlan, 1987; McLachlan and Peel, 2000). McLachlan (1987) described the process of bootstrapping the LRT statistics in a mixture of normal distributions. A more recent study by Karlis and Xekalaki (1999) introduced the use of parametric bootstrap methods sequentially to identify the number of latent classes in a mixed Poisson model. They claimed that when there are a sequence of candidate numbers of latent classes, $k, k+1, k+2, \cdots$, one should start the LRT with the lowest consecutive pair k and $k+1$. If the data offer enough evidence to reject the null hypothesis that there are k latent classes, one will perform a LRT for the next consecutive pair $k+1$ and $k+2$. The above process will be repeated until the first time one fails to reject the null hypothesis. However, Karlis and Xekalaki (1999) did not adjust the type I

error of these multiple tests since they used the same desired Type I error of 0.05 for all the tests.

Research also exists on the power of the bootstrap LRT in mixture models (McLachlan and Peel, 2000; Nylund et al., 2007; Tekle et al., 2016). Nylund et al. (2007) shows that the bootstrap likelihood ratio test is more consistent at identifying the correct number of latent classes than the information-based criteria BIC. They generated data under various scenarios, and they found that even at its worst, the bootstrap likelihood ratio test successfully detects the true number of latent classes around 49% of the time. However, in the same scenario with an eight-item categorical outcome, BIC could not select the correct model. Moreover, the bootstrap likelihood ratio test has the benefit of being consistently reliable regardless of sample size. However, after reviewing 38 articles, Tein et al. (2013) found that the bootstrap LRT test is used less for model selection in mixture models, compared with BIC, AIC, and other model selection methods.

The bootstrap is primarily criticized for its computational burden. Nylund et al. (2007) found that when applying the bootstrap LRT, the computation time increased 5 to 35 times in their examples, compared with using BIC to select the number of latent classes. To mitigate the amount of computation time, Nylund et al. (2007) set an early stopping rule for the bootstrap LRT, such that if the first n_b bootstraps demonstrated strong enough evidence to reject or fail to reject the null hypothesis, one could stop drawing further bootstrap samples. However, this early stopping rule is controversial, since it was specific to the scenario generated in Nylund et al. (2007). Moreover, when the true p-value based on infinitely many replications is around 0.05, the probability of agreement between their early stopping rule and the infinite replication procedure is less than 75%.

We propose to select the number of latent classes in JMLC with the LRT, using the parametric bootstrap to capture the null distribution of the LRT statistics. Our research is motivated by the study of acute graft-versus-host disease (aGVHD) at the University of Michigan. aGVHD is an inflammatory disease caused by a reaction between the donors' and recipients' tissues among patients who have received bone marrow transplantation. Patients' times-to-aGVHD were recorded, together with their biomarker observations up to the times-to-aGVHD or administrative censoring. There are clearly at least two latent classes of patients: the first being patients who will never develop aGVHD (aGVHD-free), and the second being patients who will experience aGVHD within 100 days of transplant. However, current research is unsure of how many subgroups there are among the patients who will experience aGVHD within 100 days.

In Section 5.2, we briefly introduce JMLC and describe how to select the number of latent classes in JMLC with bootstrap LRT, together with early stopping methods which adaptively reduce the number of bootstraps. Section 5.3 presents the simulation results of bootstrap LRT with simulated aGVHD data. Concluding remarks and discussion are given in Section 5.4.

5.2 Methods

In this section we describe how to apply JMLC to compute the MLEs of the longitudinal process and the time-to-event process, and then specify using the parametric bootstrap to select the number of latent classes in JMLC, followed by discussions on adaptively reducing the number of bootstraps.

5.2.1 JMLC model specification and model fitting

Let $\boldsymbol{z_i} = (z_{i1}, z_{i2}, \cdots, z_{iJ})'$ represent the unobserved indicator vector of subject i's latent class membership, where $z_{ih} = 1$ if subject i belongs to latent class $h = 1, 2, \cdots, J$, and 0 otherwise, and J is the number of latent classes. Let $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \cdots, \pi_{iJ})'$ be the corresponding probabilities of latent class membership, and $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_J)'$ be the marginal probabilities of subjects belonging to each aGVHD class, with the restriction that $\sum_{h=1}^{J} \pi_h = 1$.

For the longitudinal biomarker process, we define $\boldsymbol{Y}_i = (Y_i(t_1), Y_i(t_2), \cdots, Y_i(t_{n_i}))$ as the biomarker history of subject i at times $(t_1, t_2, \cdots, t_{n_i})$, where n_i is the total number of biomarker observations for subject i, for $i = 1, 2, \cdots, n$. We specify that patients from the same latent class share the same mean biomarker trajectory, with individual-specific random effects $\boldsymbol{b_i}$ reflecting the deviation of an individual's biomarker pattern from the mean of their latent class. The measurement error $\boldsymbol{e_i} = (e_i(t_1), e_i(t_2), \cdots, e_i(t_{n_i}))$ of biomarkers introduces the random noise in biomarker measurement. We assume the joint of the observed biomarkers, $\boldsymbol{Y_i}$, random effects, $\boldsymbol{b_i}$, and the measurement error, $\boldsymbol{e_i}$, $\boldsymbol{B_i}' = (\boldsymbol{Y_i}, \boldsymbol{b_i}, \boldsymbol{e_i})$ given $z_{ih} = 1$ follows a multivariate normal distribution, i.e.,

$$\boldsymbol{B_i}|z_{ih} = 1 \sim \mathcal{MVN}\left(\begin{pmatrix} X_i\boldsymbol{\beta^{(h)}} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} Z_i D Z_i^T + \sigma^2 \boldsymbol{I_{n_i}} & Z_i D & \sigma^2 \boldsymbol{I_{n_i}} \\ D Z_i^T & D & \mathbf{0} \\ \sigma^2 \boldsymbol{I_{n_i}} & \mathbf{0} & \sigma^2 \boldsymbol{I_{n_i}} \end{pmatrix}\right) \tag{5.1}$$

with density function $f_h(\boldsymbol{B_i})$, where X_i is the design matrix of function of time, $\boldsymbol{\beta^{(h)}}$ is the corresponding parameters of the mean biomarker trajectory in the latent class h, Z_i is the design matrix of random effects that can be any subset of X_i, D is the covariance matrix of the random effects that is constant among all different latent classes, and σ^2 is the common variance of each element of $\boldsymbol{e_i}$. Let

$\boldsymbol{\omega} = (\boldsymbol{\beta^{(1)}}, \boldsymbol{\beta^{(2)}}, \boldsymbol{\beta^{(3)}}, \boldsymbol{\beta^{(4)}}, D, \sigma^2)$ denote all the parameters involved in the longitudinal process.

For the time-to-event process, let T_i denote the observed event time for subject i, which is the minimum of the time-to-event for subject i, T_i^*, and last follow-up time S_i. We also define δ_i as the indicator of whether subject i experiences the event ($\delta_i = 1$) or is censored ($\delta_i = 0$). For simplicity, we assume the time-to-event follows $g_h(T_i, \delta_i|\boldsymbol{\lambda}_h)$ distribution, for $h = 1, 2, \cdots, J$, where $\boldsymbol{\lambda}_h$ are the parameters involved in the time-to-event process in latent class h.

Let $\boldsymbol{\xi} = (\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{\lambda})$ represent the complete parameter set; and

$$((\boldsymbol{Y_1}, T_1, \delta_1), (\boldsymbol{Y_2}, T_2, \delta_2), \cdots, (\boldsymbol{Y_n}, T_n, \delta_n))$$

are the n independent observed data containing both biomarker and time-to-event observations. So the data likelihood function is:

$$\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{B}, T, \delta) \;=\; \prod_{i=1}^{n} \sum_{h=1}^{J} [\pi_h f_h(\boldsymbol{B_i}|\boldsymbol{\omega}) g_h(T_i, \delta_i|\boldsymbol{\lambda})] \tag{5.2}$$

with corresponding observed data log-likelihood:

$$l(\boldsymbol{\xi}|\boldsymbol{B}, T, \delta) \;=\; \sum_{i=1}^{n} log\{\sum_{h=1}^{J} [\pi_h f_h(\boldsymbol{B_i}|\boldsymbol{\omega}) g_h(T_i, \delta_i|\boldsymbol{\lambda})]\} \tag{5.3}$$

The log-likelihood in Equation 5.3 based on observed data containing a summation in the *log* function and as such it is inconvenient to maximize. We then introduce the unobserved label of latent classes and construct the likelihood function based on

the complete data $(\boldsymbol{B_i}, T_i, \delta_i, \boldsymbol{z_i})$. The complete data likelihood function is:

$$\mathcal{L}_c(\boldsymbol{\xi}|\boldsymbol{B}, T, \delta, \boldsymbol{z}) \;=\; \prod_{h=1}^{J}[\pi_h f_h(\boldsymbol{B_i}|\boldsymbol{\omega}) g_h(T_i, \delta_i|\boldsymbol{\lambda})]^{z_{ih}} \tag{5.4}$$

with corresponding log-likelihood:

$$l_c(\boldsymbol{\xi}|\boldsymbol{B}, T, \delta, \boldsymbol{z}) \;= l_1(\boldsymbol{\pi}|\boldsymbol{B}, T, \delta, \boldsymbol{z}) + l_2(\boldsymbol{\omega}|\boldsymbol{B}, T, \delta, \boldsymbol{z}) + l_3(\boldsymbol{\zeta}|\boldsymbol{B}, T, \delta, \boldsymbol{z}) \tag{5.5}$$

where $l_1(\boldsymbol{\pi}|\boldsymbol{B}, T, \delta, \boldsymbol{z}) = \sum_{i=1}^{n}\sum_{h=1}^{J} z_{ih} log(\pi_h)$, $l_2(\boldsymbol{\omega}|\boldsymbol{B}, T, \delta, \boldsymbol{z}) = \sum_{i=1}^{n}\sum_{h=1}^{J} z_{ih} log f_h(\boldsymbol{B_i}|\boldsymbol{\omega})$, and $l_3(\boldsymbol{\lambda}|\boldsymbol{B}, T, \delta, \boldsymbol{z}) = \sum_{i=1}^{n}\sum_{h=1}^{J} z_{ih} log[g_h(T_i, \delta_i|\boldsymbol{\lambda})]$, which are three separable parts corresponding to $\boldsymbol{\pi}, \boldsymbol{\omega}$ and $\boldsymbol{\lambda}$. The Expectation-Maximization (EM) algorithm is used to find the MLEs of parameters. Using the EM algorithm to maximize the aforementioned log-likelihood is not of key interest and the details are omitted here. Please refer to Chapter II for further details.

5.2.2 Parametric bootstrap in JMLC

One prerequisite of using JMLC is specification of the number of latent classes J. We postulate that the samples are from a mixture distribution with either J_0 or J_1 components, where J_0 and J_1 are known integers with $J_0 < J_1$. By constructing the corresponding log-likelihood functions as in Equation 5.5 under the null and alternative hypotheses, $\mathcal{H}_0 : J = J_0$ vs. $\mathcal{H}_1 : J = J_1$ respectively, we compute the MLEs $\hat{\boldsymbol{\xi}}^{J_0} = \arg\max_{\Theta^{J_0}} l_0(\boldsymbol{\xi}^{J_0}|\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\delta})$ and $\hat{\boldsymbol{\xi}}^{J_1} = \arg\max_{\Theta^{J_1}} l_1(\boldsymbol{\xi}^{J_1}|\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\delta})$ through the EM algorithm as described in Section 5.2.1, where Θ^{J_0} is the parameter space under the null hypothesis, which is nested within Θ^{J_1}, the parameter space under the alternative hypothesis. Based on these MLEs, we calculate the observed LRT statistic $\mathcal{LR}_{obs} = 2(l_1(\hat{\boldsymbol{\xi}}^{J_1}|\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\delta}) - l_0(\hat{\boldsymbol{\xi}^{J_0}}|\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\delta}))$ based on Equation 5.3. As discussed in

the previous section, the exact asymptotic reference distribution of the LRT statistic is difficult to derive. Thus we will use the parametric bootstrap to determine the empirical distribution of the LRT statistic.

Specifically, we will simulate data $(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\delta})_{sim}$ under the null hypothesis from a mixture distribution of J_0 components with parameters $\hat{\boldsymbol{\xi}}^{J_0}$ in each bootstrap, with a sample size equal to that in the original data. In the k^{th} bootstrap, we fit two individual models with J_0 and J_1 components, and obtain the corresponding MLEs $\hat{\boldsymbol{\xi}}^{J_0}_{sim}$ and $\hat{\boldsymbol{\xi}}^{J_0}_{sim}$, and the LRT statistic based on the simulated data $\mathcal{LR}^k_{sim} = 2(l_1(\hat{\boldsymbol{\xi}}^{J_1}_{sim}|(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\delta})_{sim}) - l_0(\hat{\boldsymbol{\xi}}^{J_0}_{sim}|(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\delta})_{sim})$.

We repeat the above process B times, and then compare the observed LRT statistic $\mathcal{LR}_{obs}$ to its empirical null distribution $\mathcal{LR}^k_{sim}$, for $k = 1, 2, \cdots, B$. For a given α level, we will reject the null hypothesis when $\sum_{k=1}^{B} \mathcal{I}(\mathcal{LR}_{obs} > \mathcal{LR}^k_{sim})/B > 1 - \alpha$. In other words, the p-value for this bootstrap LRT is:

$$p = \sum_{k=1}^{B} \mathcal{I}(\mathcal{LR}_{obs} \leq \mathcal{LR}^k_{sim})/B \tag{5.6}$$

In order to capture the tail of this empirical null distribution, we need to choose an adequately large value for B. In practice, B is usually defined at $1,000$, $2,000$, or $10,000$ (Efron and Tibshirani, 1993; MacKinnon, 2009; Hesterberg, 2015).

It is well-known that the results of the EM algorithm depend on the initial parameter values (Karlis and Xekalaki, 2003; Biernacki et al., 2003). Therefore, to eliminate the effect of initial values, we suggest that after choosing the ideal initial parameter values for both JMLC with J_0 and J_1 components in the observed data, we save the two starting values and use them repeatedly in the bootstrap samples.

5.2.3 Adaptively reducing the number of bootstraps

In the previous section, we stated that we needed to choose a sufficiently large number of bootstrap samples to detect the behavior of LRT at the tail of its empirical distribution. We chose to use $B = 1,000$ in our methods. Obviously, if we could use fewer bootstraps but draw the same conclusion as with 1,000 bootstraps, we could save computational time. Moreover, we need to control the overall Type I error with fewer bootstraps to remain at the desired $\alpha-$level, as well as maintain the power of the LRT.

To obtain reliable results based on a smaller value of B, Davidson and MacKinnon (2000) introduced an iterative procedure. They proposed to start with a small B_0, and then evaluate whether the p-value obtained with these B_0 bootstraps is evidence to reject the null hypothesis. If the p-value based on these B_0 bootstraps was $\hat{Pr}(B_0) < \alpha$, then one would further test $\mathcal{H}_0 : Pr(B_0) < \alpha$ vs. $\mathcal{H}_1 : Pr(B_0) \geq \alpha$, at a pre-specified significance level β, which was chosen to be small, say 0.001. This test was done through a binomial approximation for the number of bootstrap samples that had statistics larger than our observed value. This significance level β could be viewed as a tuning parameter, which controlled how much we could believe in the results based on these B_0 bootstraps, and controlled the total number of bootstraps needed. This process was continued with increasing B_0 until the first time we failed to reject the null hypothesis or until we reached the upper bound of the number of bootstraps.

In another study, Nylund et al. (2007) discussed a sequential early stopping rule for bootstrap LRTs. The basic idea was to choose an adaptive α-level for each number

of bootstraps. If the p-value calculated based on the first 49 bootstraps was exactly zero, Nylund et al. (2007) suggested stopping and rejecting the null hypothesis. If the p-value based on the first 49 bootstraps was greater than zero, one continued the bootstrap process and rejected the null hypothesis with 78 bootstraps if the p-value based on these 78 bootstraps was no more than 1/78. Together with this rule, Nylund et al. also introduced an early stopping rule when the data showed strong evidence that we would fail to reject the null. More specifically, if the p-values based on the first two or three bootstraps were ≥ 1 and $\geq 2/3$, respectively, one would stop and conclude that we had failed to reject the null hypothesis. This sequential rule was justified by the concordance of its conclusion with the conclusion drawn with infinite bootstraps. The concordance probability showed that when the p-value based on an infinite number of bootstraps was quite different from the targeted $\alpha-$ level, this early stopping rule supported above 95% concordance with the infinite bootstraps. However, when the p-value based on infinite bootstraps was around the targeted α-level, the concordance probability was only around 65%.

The early stopping method introduced by Davidson and MacKinnon (2000) is a testing process that requires one to evaluate the results at each candidate value of B_0. In contrast, the method proposed by Nylund et al. (2007) is a rule that is designed before running bootstraps. Moreover, in the iterative method by Davidson and MacKinnon (2000), the significance level β at each candidate number of bootstraps should be chosen with caution. With an overly small significance level, the early stopping rule is irrelevant because there is very little chance to reject the null hypothesis that $\mathcal{H}_0 : Pr(B_0) < \alpha_0$. In contrast, an overly large significance level might lead to early stopping with the wrong conclusion. However, in Nylund et al.'s method, early stopping based on only two or three bootstraps is questionable. The concordance probability that Nylund et al. used to justify their early stopping rule only reflects

the percentage of making wrong conclusions, but it does not distinguish the chance to make a false positive or negative conclusion. Thus, we propose a new sequential early stopping rule that maintains a desired Type I error rate; we will examine the power of our approach via simulation.

We consider two types of early stopping. The first is stopping early in favor of not rejecting the null hypothesis. For example, if we run $B = 1000$ bootstraps at $\alpha = 0.05$ and observe more than 50 bootstrap LRT statistics larger than the observed LRT statistic based on original data, we would fail to reject the null hypothesis. In that case, if we observe more than 50 bootstraps with larger LRT statistics within the first B_m ($B_m \leq B$) bootstraps, we can stop after B_m bootstraps, and conclude that we have failed to reject the null hypothesis. This type of early stopping will not affect the Type I error or power of the bootstrap LRT, compared with B bootstraps.

The other type of early stopping is in favor of rejecting the null hypothesis. Similar to Nylund et al., we set a sequential stopping rule that at the pre-chosen numbers of bootstrap samples B_1, B_2, $\cdots$, we assess the numbers of bootstraps with larger LRT statistics, and compare them to pre-specified upper thresholds $U_1, U_2, \cdots$. More specifically, if we observe F_1 out of B_1 bootstraps with larger LRT statistics, with $F_1 = \sum_{k=1}^{B_1} I(LRT_{sim}^k \leq LRT_{obs}) \leq U_1$, we will stop and conclude that we have rejected the null hypothesis. If we fail to reject the null hypothesis with the first B_1 bootstraps, we will continue with more bootstraps. At the mth decision point, which occurs with B_m bootstraps, we will calculate the number of bootstraps with larger LRT statistics than the original one and compare this number F_m with the threshold U_m. As long as the number of larger LRT statistics is between U_m and αB, we will continue the bootstrap process until we complete B bootstraps, and draw the final conclusion given the complete B bootstraps.

The sequence of the early stopping rule values $((B_1, U_1), (B_2, U_2), \cdots)$ is analogous to an alpha-spending function in clinical trials (Demets and Lan, 1994). The thresholds U_m should be chosen to control the overall Type I error and also minimize the loss of power. For example, if we set the targeted α-level at 0.05 for 1000 bootstraps, and we observe 2 bootstrap LRT statistics larger than the observed LRT statistic of the original data within the first 200 bootstraps, it is reasonable for us to believe that we will end up rejecting the null hypothesis with 1000 bootstraps. However, we could also make an incorrect decision. If we observe more than 48 larger bootstrap LRT statistics within the next 800 bootstraps, we might falsely reject the null hypothesis with the first 200 bootstraps.

With a targeted overall α-level at 0.05 and an upper bound of B at 1000, we propose two rules $((B_1, U_1), (B_2, U_2), \cdots)$ for comparison:

- Rule 1: (200, 8), (400, 17), (600, 22), (800, 30), (1000, 40)
- Rule 2: (200, 6), (400, 14), (600, 26), (800, 35), (1000, 47)

These two rules make decisions in increments of 200 bootstraps, so that at most five assessments are made. Both of these rules control the overall Type I error to be no more than 0.05. The calculation details of Type I error can be found in Appendix A.1.

Unlike the Type I error, evaluating the above two candidate rules with regards to power is not straightforward. This is because the distribution of p-values under the alternative hypothesis is a function of both sample size and the effect size in the alternative hypothesis, making it difficult to derive in closed form. Given that the p-value under the null hypothesis follows a uniform distribution, which is a special case of a Beta distribution, we hypothesize a theoretical Beta (1,31.4) distribution for the p-value under the alternative hypothesis. If a variable X follows a Beta (1,31.4) distri-

bution, the probability of X less than 0.05 is around 0.8. In other words, if the p-value for one test under the alternative hypothesis follows a Beta (1,31.4) distribution, then the power of this test is around 0.8. When applying our two candidate rules, the losses of power for these rules are 0.0136 and 0.0069, respectively. It is because in order to reject the null hypothesis, both of the two rules require a more restricted p-value than 0.05. It is clear that the loss of power is controlled by the largest value of U_m, so that candidate rule 1 has the larger loss of power. However, compared with the targeted power 0.8, both of these losses of power are relatively trivial. Moreover, the choice of rules should also take the computational time into consideration. Rule 1 is more likely to terminate the bootstrap process earlier because it is more tolerant in early stages. We will evaluate the power of the two stopping rules in the simulation section.

In practice, we will apply the early stopping rules in favor of both the null and the alternative hypothesis simultaneously. As demonstrated in Figure 5.1, the number of bootstraps with larger LRT statistics than that observed is calculated at every 200 bootstraps. If that number falls in the "rejecting zone", one will stop there and reject the null hypothesis. On the other hand, the early stopping in favoring of the null hypothesis is assessed whenever a new bootstrap is available after the first 50 bootstraps. As soon as the number of bootstraps with larger LRT statistics falls into the "fail to reject zone", one will stop and fail to reject the null hypothesis. Otherwise, if that number falls between the two threshold lines, one will draw an additional 200 bootstraps and apply the next decision rule.

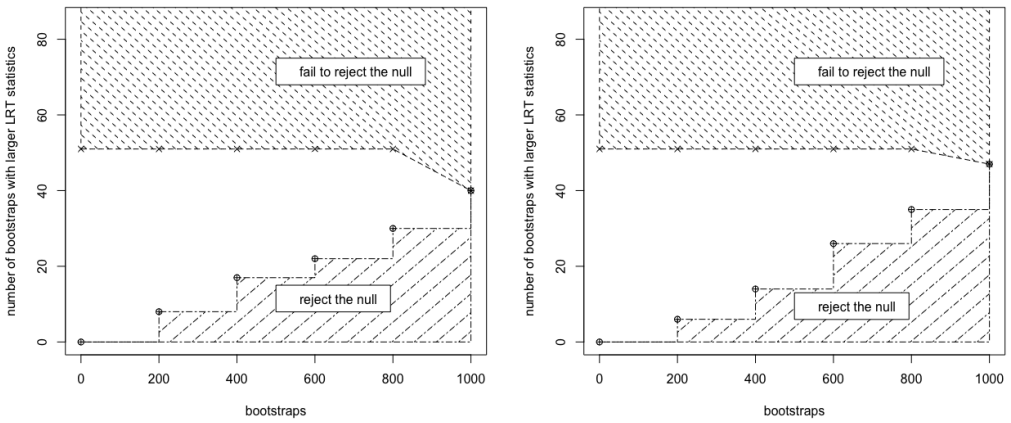

Figure 5.1: Diagram of early stopping rules adaptively reducing the number of bootstraps (left: rule 1; right: rule 2)

5.3 Simulation and Result

5.3.1 Comparing Bootstrap LRT and BIC

The purpose of this simulation is to examine the Type I error and the power of bootstrap LRT in testing one pair of candidate numbers of latent classes. The BICs of the two models with different numbers of latent classes are also calculated. The accuracy rate of model selection based on BIC, which is defined as the proportion of times that the true model has a lower BIC, is calculated. We use BIC as a benchmark, and evaluate the performance of bootstrap LRT relative to BIC.

Here we focus on testing one pair of candidate numbers of latent classes: $\mathcal{H}_0 : J_0 = 3$ vs. $\mathcal{H}_0 : J_1 = 4$. Data are simulated under both the null and alternative hypotheses in order to examine the Type I error and the power of the parametric bootstrap LRT, respectively. One thousand simulations are run under each hypothesis.

In each simulation, a sample of 180 patients is generated. Given the latent class, each patient's biomarker observations are generated from a linear mixed model, with a

class-specified mean biomarker trajectory and a common variance-covariance matrix for random effects shared across latent classes. The times-to-aGVHD follow Weibull distributions with class-specified shape and scale parameters, except for the aGVHD-free patients. Biomarker values of each patient are collected weekly right after the transplant, until the earliest of onset of aGVHD or day 100. Since we assume that the baseline biomarker value is non-informative for time-to-aGVHD, we require that each subject should have at least two observations. As patients who develop aGVHD within one week after the transplantation will be removed from the study, the sample size might be slightly less than 180. The patients' times-to-aGVHD are recorded together with the biomarker values. Two JMLC models are fitted under the null hypothesis $J_0 = 3$ and the alternative $J_1 = 4$, respectively. The observed LRT statistic LRT_{obs} is calculated and recorded. One thousand parametric bootstrap samples are generated from the null distribution ($J_0 = 3$) as described in Section 5.2.2. We set the desired Type I error rate at $\alpha = 0.05$, and reject the null hypothesis in each simulation if $\sum_{k=1}^{1000} \mathcal{I}(\mathcal{LRT}_{obs} > \mathcal{LRT}^k_{simu})/1000 > 1 - \alpha$.

First we simulate data under the null hypothesis $\mathcal{H}_0 : J_0 = 3$, and test against the alternative hypothesis of $\mathcal{H}_1 : J_1 = 4$. Figure 5.2 presents a sample of data we simulated; the left panel shows the observed biomarker values overall, while the right panel highlights the values by latent group membership. With the 1,000 simulations under this scenario, we calculate the probability of rejecting the null hypothesis, which is the Type I error of the parametric bootstrap LRT. We also calculate the proportion of times that the joint model with three latent classes has a lower BIC than the joint model with four latent classes.

In these 1,000 simulations, 54 simulations reject the null hypothesis according to the bootstrap LRT, so the Type I error of bootstrap LRT is 0.054. However, 710 sim-

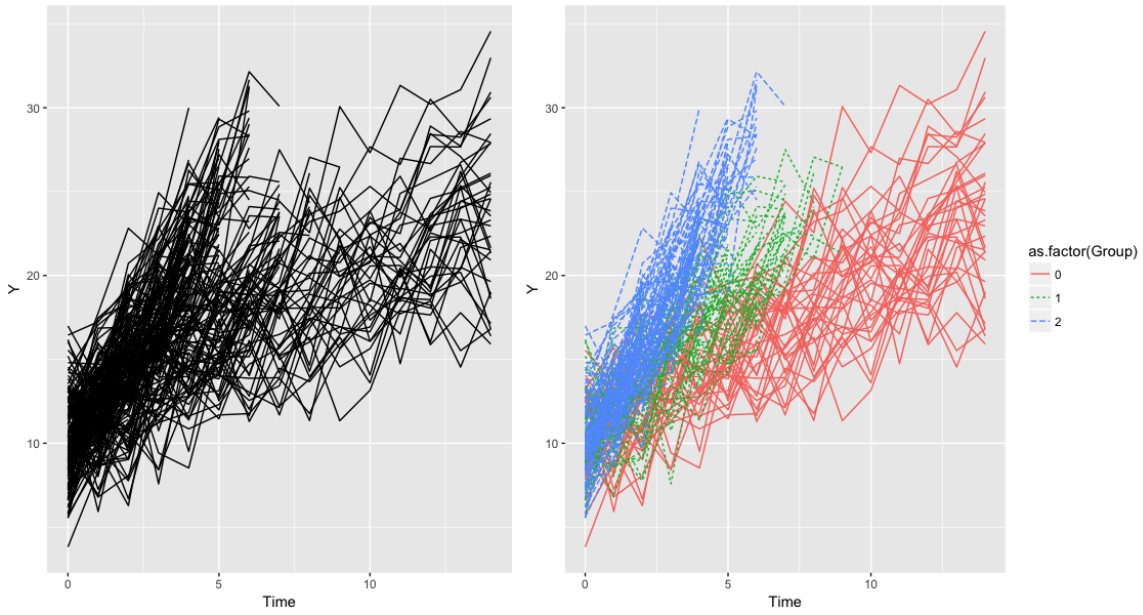

Figure 5.2: Biomarker observations of patients from three latent classes (left: overall; right: by latent groups)

ulations show that joint model with four latent classes is preferred in terms of lower BIC; the accuracy rate for BIC is therefore 29%. The bootstrap LRT controls the Type I error well, while BIC shows a low accuracy rate when the data are simulated from the null distribution.

In order to evaluate the power of the parametric bootstrap LRT, data are simulated under the alternative hypothesis $\mathcal{H}_1 : J_1 = 4$. Figure 5.3 presents a sample biomarker data of patients from four latent groups. Though the four latent classes are highlighted in the right panel, it is hard to tell whether there are three or more latent classes. With the 1,000 simulations under this setting, we calculate the probability of rejecting the null hypothesis, which is the power of the parametric bootstrap LRT. We also calculate the proportion of times that the joint model with four latent classes has a lower BIC than the joint model with three latent classes.

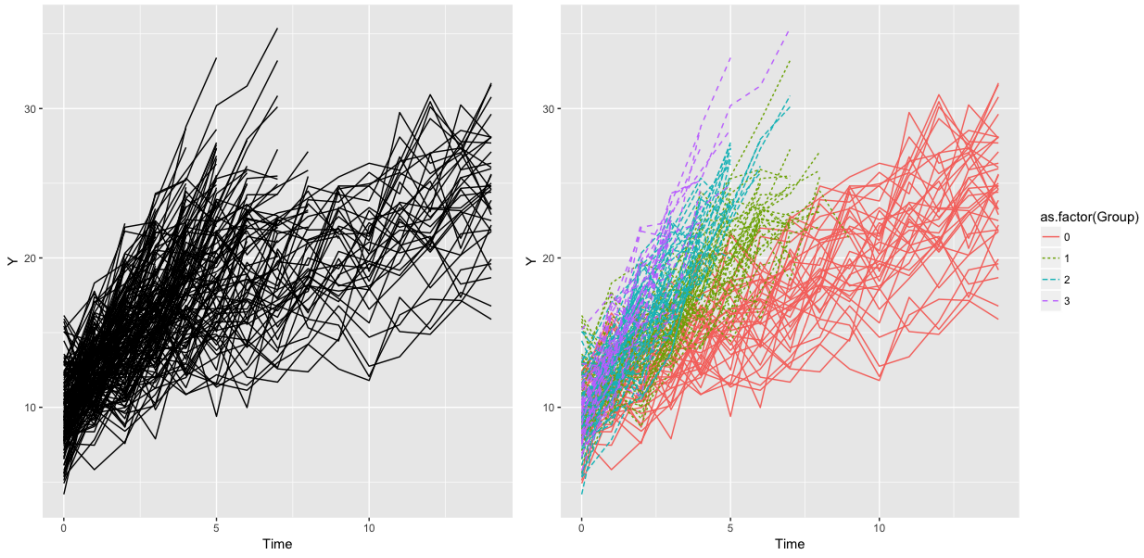

Figure 5.3: Biomarker observations of patients from four latent classes (left: overall; right: by latent groups)

Out of those 1,000 simulations, 727 simulations reject the null hypothesis according to the bootstrap LRT, so the power of bootstrap LRT is 0.727. Out of these 1000 simulations, 992 simulations show that the joint model with four latent classes are preferred in terms of lower BIC. The accuracy rate for BIC is 99.2%.

Compared the results listed above, we found that BIC tends to favor more complex models while bootstrap LRT achieves good power while controls the Type I error.

5.3.2 Type I Error and Power of Bootstrap LRT with Adaptive Reduction in the Number of Bootstraps

In this section we will apply the two candidate early stopping rules introduced in Section 5.2.3, together with the early stopping rule in favor of the null hypothesis, to the simulation settings examined in Section 5.3.1. We will first evaluate the Type I error of the two rules, followed by the power and the computational time of the two

rules.

Recall that the early stopping rule in favor of the null hypothesis states that we will stop generating bootstrap samples within the first B_x ($B_x \leq B$) bootstraps if more than 50 of the B_x samples have larger LRT statistics than the observed LRT statistic, we will conclude that we have failed to reject the null hypothesis. Among our 1000 simulations, the average number of bootstraps was 212. This means with this early stopping rule in favor of the null hypothesis, we reduced our computational times by 80% when the null hypothesis is true.

When applying rule 1, 51 out of 1000 simulations reject the null hypothesis, so the Type I error for rule 1 is 0.051; when applying rule 2, 53 out of 1000 simulations reject the null hypothesis, so the Type I error for rule 2 is 0.053. The two rules are more conservative requiring no more than 40 or 47 bootstraps with larger LRT statistics, compared with standard rule that requires no more than 50 bootstraps out of 1000. When applying rule 1, the average number of bootstraps is reduced to 173; and the average number of bootstraps is 178 when applying rule 2.

Given the data are simulated under the alternative hypothesis, the average number of bootstraps among these 1000 simulations is 839 when apply the early stopping rule in favor of the null hypothesis. This means with this early stopping rule, we reduce our computational time by approximately 16%.

When applying rule 1, 730 out of 1000 simulations reject the null hypothesis, so the power for rule 1 is 0.730; when applying rule 2, 728 out of 1000 simulations reject the null hypothesis, so the power for rule 2 is 0.728. The difference between these values is due to 23 simulations, summarized in Table 5.1.

For example, based on results of all 1,000 bootstraps in simulations 2, 9, and 10,

Table 5.1: 23 simulation results with contradictory conclusions based on early stopping rules and 1000 bootstraps

ID	Number of larger LRT statistics within first N* bootstraps					Rule 1	Rule 2	1000 Bootstraps
	200	400	600	800	1000			
1	7	18	27	40	51	1	0	0
2	6	17	32	45	51	1	1	0
3	10	18	23	33	44	0	1	1
4	3	13	31	38	51	1	1	0
5	10	19	28	42	48	0	0	1
6	10	20	30	38	41	0	1	1
7	11	19	29	36	43	0	1	1
8	7	16	28	42	51	1	0	0
9	4	15	27	34	51	1	1	0
10	6	25	35	45	51	1	1	0
11	12	23	31	36	48	0	0	1
12	8	24	38	47	51	1	0	0
13	12	25	29	37	48	0	0	1
14	8	30	45	51	51	1	0	0
15	8	28	46	51	51	1	0	0
16	13	22	31	39	48	0	0	1
17	7	20	34	46	51	1	0	0
18	12	21	27	36	44	0	1	1
19	7	19	30	43	51	1	0	0
20	9	15	25	46	51	1	1	0
21	7	17	34	44	51	1	0	0
22	10	22	31	40	43	0	1	1
23	13	19	27	37	45	0	1	1

one would have failed to reject the null hypothesis. However, when applying the two early stopping rules, one would stop early and reject the null hypothesis. In contrast, the 1,000 bootstraps in simulations 5 and 11 support the alternative hypothesis, while with these two early stopping rules, one would fail to reject the null hypothesis.

When applying rule 1, the average number of bootstraps is reduced to 264, while the average number of bootstraps is 280 when applying rule 2. Both of the two rules obviously reduce computational times, but are able to maintain sufficient power.

5.4 Discussion

In this project, we described the procedure of the bootstrap LRT in the selection of the number of latent classes in JMLC, and explore the performance of bootstrap LRT in terms of Type I error and testing power. BIC, the current standard of model selection in the number of latent classes, tends to select a more complex model with more latent classes. Compared with BIC, bootstrap LRT controls well the Type I error, and maintain sufficient power.

BIC adds a penalty term for the number of parameters in the model, k, to the negative log-likelihood. Thus, the difference of BICs between two nested models equals the LRT statistic, minus a value c which only depends on the difference of the numbers of parameters in the two models and the sample size n, as shown in Equation 5.7:

$$BIC(M_0) - BIC(M_1) = 2[log(l_{M_1}) - log(l_{M_0})] - log(n)[k_{M_1} - k_{M_0}] \tag{5.7}$$

Bootstrap method find an empirical threshold, l^*, for LRT statistics that if the observed LRT statistic is above l^*, one will reject the null hypothesis. According to Equation 5.7, the bootstrap process is equivalent to selecting the richer model M_1 if the difference of BICs between the two model M_0 and M_1 is larger than $l^* + c$. Thus, one model is selected if the difference of BICs between two models exceeds a threshold. However, in practice, we do not know this threshold.

In order to make a robust selection of the number of latent classes, one needs to choose a sufficiently large B, which leads to the computational burden of bootstrap LRT. In our project, we propose two candidate early stopping rules in favor of the alternative hypothesis that can adaptively reduce the number of bootstraps. These

two rules examine the results of every 200 bootstraps, and will terminate the bootstrap process if there is already enough evidence to reject the null hypothesis given available bootstraps. The simulation results suggest that these two rules will not cause dramatic power loss, but save the computation times significantly.

In contrast to the early stopping rule in favor of the null hypothesis, which assesses the bootstraps results whenever a new bootstrap is generated, these two rules only examine the bootstrap results when every additional 200 bootstraps are available. This is because the early stopping rule in favor of the null hypothesis does not change the Type I error or power. The two early stopping rules in favor of the alternative hypothesis, however, might change the Type I error and the power of the bootstrap process. Without careful correction of the Type I error at each assessment, multiple testing will inflate the overall Type I error. The upper bounds used with 1,000 bootstraps in our two rules are set to be 40 and 47, respectively, to control for the overall Type I error. One could always change these rules, for example, by assessing the bootstrap results whenever an additional 100 bootstraps are available. In an extreme example, Nylund et al. (2007) drew conclusions using only the first two or three bootstraps. In our project, we propose to assess the bootstrap results with every 200 bootstraps to balance the computational time and the reliability of the results. Conclusions drawn with too few bootstraps are questionable due to randomness, while too many bootstraps will harm our goal of saving computational time.

The two early stopping rules we proposed in Section 5.2.3 are two examples that reach a good balance between adequate power and manageable computation time. More specifically, we do not lose too much power due to multiple assessments, but can save considerable computation time. With a targeted overall Type I error and a fixed maximum number of bootstraps B, one could construct his/her own early stop-

ping rules in favor of the alternative hypothesis. One could update his/her decision when every additional B_m bootstraps are available, and evaluate the results with a customized upper limit list $(U_1, U_2, \cdots)$.

In practice, we might not have one specific pair of numbers of latent classes, J_0 and J_1, to test, but rather an ordered sequence of candidate values $J_0, J_1, \cdots, J_{max}$, where J_{max} is the maximum possible number of latent classes. Schlattmann and Böhning (1993) described a backward selection process, where they failed to reject the first hypothesis $\mathcal{H}_0 : J = 4$ vs. $\mathcal{H}_1 : J = 5$ but rejected the second one $\mathcal{H}_0 : J = 3$ vs. $\mathcal{H}_1 : J = 4$, both at an α-level at 0.05. Karlis and Xekalaki (1999), in contrast, proposed a forward selection process. Here we describe how to select the number of latent classes through a sequential testing process, while correcting for overall Type I error.

If there is no prior information on the number of latent classes, we usually start with the assumption of homogeneous samples ($J_0 = 1$), testing this hypothesis against a mixture distribution of two components ($J_1 = 2$). If there is enough evidence to reject the null hypothesis, we would move on to test two components against three, then three against four, and so forth. We would stop the first time we fail to reject $\mathcal{H}_0 : J_0$ and conclude that there are J_0 latent classes in the population. In our aGVHD setting, patients are from at least two latent classes, the aGVHD-free class and aGVHD class, so we would start with $\mathcal{H}_0 : J_0 = 2$.

In this sequential testing process, we need to control the overall Type I error. Many sequential testing methods have been proposed and widely applied. For example, the study by Schlattmann and Böhning (1993) discussed the limitations of Bonferroni adjustment, which are the dramatic loss of power and inconsistent esti-

mation of proportions of subgroups. Another method, the alpha-spending function proposed by Demets and Lan (1994), which is widely used in interim analysis in clinical trials.Similar work has been done in detecting the change point in proportional hazard models (Goodman et al., 2011; He et al., 2013), and we adopt their alpha-spending function.

Goodman et al. (2011) gave a brief proof of why their method can control the overall Type I error . A more detailed proof of Type I error control in aGVHD setting can be found in the Appendix A.2.

If the overall significance level is α, we will use an alpha-spending function $\alpha^*(m) = \alpha/2^{m-2}$, where $\alpha^*(m)$ is the significance level for hypothesis test: $\mathcal{H}_0 : J = m$ vs. $\mathcal{H}_1 : J = m + 1$. Thus, if we start with testing $\mathcal{H}_0 : J = 2$ vs. $\mathcal{H}_1 : J = 3$, we will use the significance level of 0.05; if we reject the null hypothesis and move on to test $\mathcal{H}_0 : J = 3$ vs. $\mathcal{H}_1 : J = 4$, we will use the new significance level of 0.025, and so forth. In other words, each hypothesis will be tested under a more conservative significance level than the previous hypothesis.

It is worth noting that one major advantage of this alpha-spending function is that its calculation does not require setting the upper bound of the number of latent classes. As long as the data present enough evidence to reject the null hypothesis, one could continue testing the next consecutive pair of the numbers of latent classes.

CHAPTER VI

Summary

In this dissertation, we build three prediction tools to dynamically predict the onset of aGVHD with longitudinal biomarkers. Our approaches have the ability to identify subgroups of various risk, and refine the aGVHD prediction whenever a new biomarker observation is available.

We have contributed to the existing literature on the application and model selection of JMLC. Our bootstrap method in selecting the number of latent classes has been proved to be more robust than the standard information-based criteria. It has been demonstrated that our method controls the Type I error well while maintaining sufficient power. We have also proposed two sequential early stopping rules, which can save around 80% of the computational time.

We have proposed a revised landmark analysis, which uses all the information up to the landmark time to identify the subgroups of aGVHD risk. In contrast to the standard landmark analysis which uses only the biomarker observation at the landmark time, our approach can alleviate the effect of the measure error and provide more efficient aGVHD prediction.

We have also explored the performance of the pattern mixture model in our settings. The pattern mixture model is easy to execute and straightforward to interpret. Simulations have indicated that the pattern mixture model controls loss of accuracy in predictions. Moreover, we have generalized the pattern mixture model by incorporating censored cases. The simulation results have demonstrated that this generalized pattern mixture model results in more accurate estimations of the marginal pattern probability, and thus achieves higher prediction accuracy compared to the complete-case analysis of early predictions.

We have discussed the future work for each project in the corresponding discussion section for each chapter. Furthermore, we are also planning to develop a user-friendly application to allow better bench-to-bedside translational statistical tools.

APPENDIX

APPENDIX A

Bootstrap Likelihood Ratio Test: Type I Error

A.1 Type I error of the adaptive early stopping rule

When the null hypothesis is true, the p-value of the LRT test follows a uniform distribution in $[0, 1]$, which is equivalent to a Beta distribution $B(1, 1)$. Under the null hypothesis, the probability of observing no more than N_1 bootstrap LRT statistics larger than the observed one among the first 200 bootstraps is:

$$\begin{aligned} Pr\left(\sum_{k=1}^{200} \mathcal{I}(\mathcal{LR}_{obs} \leq \mathcal{LR}^k_{sim}) \leq N_1 | \mathcal{H}_0\right) &= \int_0^1 \sum_{l=0}^{N_1} \binom{200}{l} p^l (1-p)^{200-l} dp \\ &= \sum_{l=0}^{N_1} \frac{200!}{l!(200-l)!} \frac{\Gamma(l+1)\Gamma(201-l)}{\Gamma(202)} \\ &= \sum_{l=0}^{N_1} 1/201 \\ &= \frac{N_1+1}{201} \end{aligned}$$

Under the null hypothesis, the probability of observing more than N_1 larger bootstrap LRT statistics with the first 200 bootstraps, but no more than N_2 within the

first 400 bootstraps is:

$$
\begin{aligned}
& Pr\left(\sum_{k=1}^{200}\mathcal{I}(\mathcal{LR}_{obs}\leq\mathcal{LR}_{sim}^{k})>N_1,\sum_{k=1}^{400}\mathcal{I}(\mathcal{LR}_{obs}\leq\mathcal{LR}_{sim}^{k})\leq N_2|\mathcal{H}_0\right)\\
= & \int_0^1\sum_{l=N_1+1}^{N_2}\binom{200}{l}p^l(1-p)^{200-l}[\sum_{m=0}^{N_2-l}\binom{200}{m}p^m(1-p)^{200-m}]dp\\
= & \sum_{l=N_1+1}^{N_2}\sum_{m=0}^{N_2-l}\binom{200}{l}\binom{200}{m}\int_0^1 p^{l+m}(1-p)^{400-l-m}dp\\
= & \sum_{l=N_1+1}^{N_2}\sum_{m=0}^{N_2-l}\frac{200!300!}{501!}\frac{(l+m)!(400-l-m)!}{l!m!(200-l)!(200-m)!}
\end{aligned}
$$

Similarly, we can derive the formula to calculate the probability of observing more than $(N_1, N_2, \cdots, N_q)$ larger bootstrap LRT statistics with the first $(B_1, B_2, \cdots, B_q)$ bootstraps, respectively, but no more than N_{q+1} within the first B_{q+1} bootstraps. With appropriate choice of $(N_1, N_2, \cdots)$ together with $(B_1, B_2, \cdots)$, one can design the early stopping rule that controls the overall Type I error.

For example, with the early stopping rule 1 introduced in Section 5.2.3, that one will stop the bootstrap LRT process and conclude that one has rejected the null hypothesis if there are no more than eight bootstraps within the first 200 bootstraps having larger LRT statistics than the observed one. If one fails to reject the null hypothesis, he will continue and revisit the bootstraps LRT statistics at the 400th bootstrap. Here we will calculate the Type I error of the early stopping rule 1: (200,8),

(400, 17), (600, 22), (800, 30), (1000, 40).

$$
\begin{aligned}
&Pr\left(\sum_{k=1}^{200}\mathcal{I}(\mathcal{LR}_{obs}\leq\mathcal{LR}_{sim}^{k})\leq 8|\mathcal{H}_0\right)=0.044776\\
&Pr\left(\sum_{k=1}^{200}\mathcal{I}(\mathcal{LR}_{obs}\leq\mathcal{LR}_{sim}^{k})>8,\sum_{k=1}^{400}\mathcal{I}(\mathcal{LR}_{obs}\leq\mathcal{LR}_{sim}^{k})\leq 17|\mathcal{H}_0\right)=0.004114\\
&Pr\left(\sum_{k=1}^{200}\mathcal{I}(\mathcal{LR}_{obs}\leq\mathcal{LR}_{sim}^{k})>8,\cdots,\sum_{k=1}^{600}\mathcal{I}(\mathcal{LR}_{obs}\leq\mathcal{LR}_{sim}^{k})\leq 22|\mathcal{H}_0\right)=0.000163\\
&Pr\left(\sum_{k=1}^{200}\mathcal{I}(\mathcal{LR}_{obs}\leq\mathcal{LR}_{sim}^{k})>8,\cdots,\sum_{k=1}^{800}\mathcal{I}(\mathcal{LR}_{obs}\leq\mathcal{LR}_{sim}^{k})\leq 30|\mathcal{H}_0\right)=0.000314\\
&Pr\left(\sum_{k=1}^{200}\mathcal{I}(\mathcal{LR}_{obs}\leq\mathcal{LR}_{sim}^{k})>8,\cdots,\sum_{k=1}^{1000}\mathcal{I}(\mathcal{LR}_{obs}\leq\mathcal{LR}_{sim}^{k})\leq 40|\mathcal{H}_0\right)=0.000601
\end{aligned}
$$

Given these calculation results, the probability of rejecting the null hypothesis when the null hypothesis is true, is 0.049968, which is controlled under 0.05.

A.2 Overall Type I error control proof

Right now we will prove that the aforementioned sequential significance levels $\alpha^*(m)$ for $m = 1, 2, \cdots$ achieve an overall type 1 error no larger than α. Starting with no further information of the aGVHD patients, in other words, $J = 2$, the probability of reject null hypothesis $\mathcal{H}_0 : J = 2$ is:

$$
\begin{aligned}
Pr(\hat{J}>2|J=2) &= Pr(\hat{J}=3,4,\cdots|J=2)\\
&= Pr(\hat{J}=3|J=2)+Pr(\hat{J}=4|J=2)+Pr(\hat{J}=5|J=2)+\cdots\\
&= \alpha(1-\frac{\alpha}{2})+\alpha\frac{\alpha}{2}(1-\frac{\alpha}{4})+\alpha\frac{\alpha}{2}\frac{\alpha}{4}(1-\frac{\alpha}{8})+\cdots\\
&= \alpha-\prod_{m=1}^{\infty}\frac{\alpha}{2^{m-1}}\\
&\leq \alpha
\end{aligned}
$$

Similarly, if the null hypothesis $J = m$ is true, then the probability of reject null hypothesis is definitely smaller than α. So the overall Type I error in the aforementioned sequential testing process is bounded by α.

BIBLIOGRAPHY

BIBLIOGRAPHY

Akogul, S. and Erisoglu, M. (2016). A comparison of information criteria in clustering based on mixture of multivariate normal distributions. *Mathematical and Computational Applications* **21,** 34.

Behar, E., Chao, N. J., Hiraki, D. D., Krishnaswamy, S., Brown, B. W., Zehnder, J. L., and Grumet, F. C. (1996). Polymorphism of adhesion molecule cd31 and its role in acute graft-versus-host disease. *New England Journal of Medicine* **334,** 286–291.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis* **41,** 561–575.

Blanche, P., ProustLima, C., Loubere, L., Berr, C., Dartigues, J., and JacqminGadda, H. (2015). Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics* **71,** 102113.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78,** 1–3.

Cancer Facts & Figures (2016). *Cancer Facts & Figures 2016.* American Cancer Society.

Center for International Blood & Marrow Transplant Research (2005). *www.cibmtr.org/Meetings/Materials/GVHDworkshop.*

Center for International Blood & Marrow Transplant Research (2015). *Summary Slices 2015.*

Champlin, R. (2003). Selection of autologous or allogeneic transplantation. *Holland-Frei Cancer Medicine* .

Chen, J. and Khalili, A. (2008). Order selection in finite mixture models with a nonsmooth penalty. *Journal of the American Statistical Association* **103,** 1674–1683.

Chen, Y. and Cutler, C. (2013). Biomarkers for acute gvhd: can we predict the unpredictable? *Bone Marrow Transplantation* **48,** 755–760.

Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B* **66,** 165–185.

Dal Maso, L., Guzzinati, S., Buzzoni, C., Capocaccia, R., Serraino, D., Caldarella, A., Dei Tos, A., Falcini, F., Autelitano, M., Masanotti, G., and Ferretti, S. (2014). Long-term survival, prevalence, and cure of cancer: a population-based estimation for 818 902 italian patients and 26 cancer types. *Annals of Oncology* **25,** 2251–2260.

Davidson, R. and MacKinnon, J. G. (2000). Bootstrap tests: How many bootstraps? *Econometric Reviews* **19,** 55–68.

Demets, D. L. and Lan, K. (1994). Interim analysis: the alpha spending function approach. *Statistics in Medicine* **13,** 1341–1352.

Efron, B. and Tibshirani, R. J. (1993). An introduction to the bootstrap: Monographs on statistics and applied probability, vol. 57. *New York and London: Chapman and Hall/CRC* .

Fang, H.-B., Wu, T. T., Rapoport, A. P., and Tan, M. (2016). Survival analysis with functional covariates for partial follow-up studies. *Statistical Methods in Medical Research* **25,** 2405–2419.

Ferrara, J. L., Levine, J. E., Reddy, P., and Holler, E. (2009). Graft-versus-host disease. *The Lancet* **373,** 1550–1561.

Fonseca, J. R. and Cardoso, M. G. (2007). Mixture-model cluster analysis using information theoretical criteria. *Intelligent Data Analysis* **11,** 155–173.

Garnett, C., Apperley, J. F., and Pavl, J. (2013). Treatment and management of graft-versus-host disease: improving response and survival. *Therapeutic Advances in Hematology* **4,** 366–378.

Goodman, M. S., Li, Y., and Tiwari, R. C. (2011). Detecting multiple change points in piecewise constant hazard functions. *Journal of Applied Statistics* **38,** 2523–2532.

Harris, A., Ferrara, J., Braun, T., Couriel, D., Choi, S., Kitko, C., Goldstein, S., Magenau, J., Paczesny, S., Pawarode, A., and Reddy, P. (2013). A combination of clinical characteristics and day 7 biomarker concentrations predicts graft-versus-host disease following hematopoietic cell transplantation from related donors. *Biology of Blood and Marrow Transplant* **19,** S138–S139.

Harris, A. C., Taylor, A., Braun, T. M., Magenau, J., and Ferrara, J. L. (2013). A biomarker-based grading system at onset of gvhd predicts nrm better than the modified glucksberg grading system. *Blood* **122,** 145–145.

He, P., Fang, L., and Su, Z. (2013). A sequential testing approach to detecting multiple change points in the proportional hazards model. *Statistics in Medicine* **32,** 1239–1245.

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

Henderson, R., Diggle, P., and Dobson, A. (2002). Identification and efficacy of longitudinal markers for survival. *Biostatistics* **3,** 33–50.

Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician* **69,** 371–386.

Hosing, C., Saliba, R., McLaughlin, P., Andersson, B., Rodriguez, M., Fayad, L., Cabanillas, F., Champlin, R., and Khouri, I. (2003). Long-term results favor allogeneic over autologous hematopoietic stem cell transplantation in patients with refractory or recurrent indolent non-hodgkins lymphoma. *Annals of Oncology* **14,** 737–744.

Jacobsohn, D. A. and Vogelsang, G. B. (2007). Acute graft versus host disease. *Orphanet Journal of Rare Diseases* **2,** 1.

Jiang, B., Elliott, M. R., Sammel, M. D., and Wang, N. (2015). Joint modeling of cross-sectional health outcomes and longitudinal predictors via mixtures of means and variances. *Biometrics* **71,** 487–497.

Karlis, D. and Xekalaki, E. (1999). On testing for the number of components in a mixed poisson model. *Annals of the Institute of Statistical Mathematics* **51,** 149–162.

Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the em algorithm for finite mixtures. *Computational Statistics & Data Analysis* **41,** 577–590.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A* pages 49–66.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics* **20,** 1350–1360.

Levine, J., Braun, T., Harris, A., Holler, E., Taylor, A., Miller, H., Magenau, J., Weisdorf, D., Ho, V., Bolaos-Meade, J., and Alousi, A. (2014). A biomarker algorithm defines onset grades of acute graft-vs-host disease with distinct non-relapse mortality. *Blood* **124,** 661–661.

Levine, J., Braun, T., Harris, A., Holler, E., Taylor, A., Miller, H., Magenau, J., Weisdorf, D., Ho, V., Bolaos-Meade, J., and Alousi, A. (2015). A prognostic score for acute graft-versus-host disease based on biomarkers: a multicentre study. *The Lancet Haematology* **2,** e21–e29.

Levine, J., Kitko, C., Yanik, G., Braun, T., Paczesny, S., Mineishi, S., Krijanovski, O., Jones, D., Cooke, K., Whitfield, J., and Hutchinson, R. (2006). Changes in TNFR1 levels in the first week post-myeloablative hsct correlate with severity and incidence of GVHD and 1y TRM. *Blood* **108,** 37–37.

Lin, H., Turnbull, B., McCulloch, C., and Slate, E. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association* **97,** 53–65.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–163.

Lindsten, F., Ohlsson, H., and Ljung, L. (2011). *Just relax and come clustering!: A convexification of k-means clustering*. Linköping University Electronic Press.

Liu, Y., Liu, L., and Zhou, J. (2015). Joint latent class model of survival and longitudinal data: An application to cpcra study. *Computational Statistics & Data Analysis* **91,** 40–50.

Lo, Y., Mendell, N. R., and Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika* pages 767–778.

MacKinnon, J. G. (2009). Bootstrap hypothesis testing. *Handbook of Computational Econometrics* **183,** 213.

Mayeux, R. (2004). Biomarkers: potential uses and limitations. *NeuroRx* **1,** 182–188.

McCutcheon, A. L. (1987). *Latent class analysis*. Sage.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Applied Statistics* pages 318–324.

McLachlan, G. J. and Peel, D. (2000). Finite mixture models.

Mehrjou, A., Hosseini, R., and Araabi, B. N. (2016). Improved bayesian information criterion for mixture model selection. *Pattern Recognition Letters* **69,** 22–27.

Naylor, S. (2003). Biomarkers: current perspectives and future prospects.

Neyman, J. and Scott, E. (1965). On the use of c (alpha) optimal tests of composite hypotheses. *Bulletin of the International Statistical Institute* **41,** 477–497.

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling* **14,** 535–569.

Olofsen, E. and Dahan, A. (2013). Using akaike's information theoretic criterion in mixed-effects modeling of pharmacokinetic data: a simulation study. *F1000Research* **2,** 2–71.

Paczesny, S., Braun, T., Vander Lugt, M., Harris, A., Fiema, B., Hernandez, J., Choi, S., Kitko, C., Magenau, J., Yanik, G., et al. (2011). A three biomarker panel at days 7 and 14 can predict development of grade II-IV acute graft-versus-host disease. *Biology of Blood and Marrow Transplantation* **17,** S167.

Paczesny, S., Choi, S., Braun, T., Kitko, C., Oleg, K., Clouthier, S. G., Weyand, A., Bickley, D., Jones, D., Whitfield, J., et al. (2007). A four protein plasma fingerprint of acute graft versus host disease (gvhd) predicts long term survival. *Blood* **110,** 38–38.

Paczesny, S., Krijanovski, O. I., Braun, T. M., Choi, S. W., Clouthier, S. G., Kuick, R., Misek, D. E., Cooke, K. R., Kitko, C. L., Weyand, A., et al. (2009). A biomarker panel for acute graft-versus-host disease. *Blood* **113,** 273–278.

Paczesny, S., Levine, J. E., Braun, T. M., and Ferrara, J. L. (2009). Plasma biomarkers in graft-versus-host disease: a new era? *Biology of Blood and Marrow Transplantation* **15,** 33–38.

Proust-Lima, C., Dartigues, J.-F., and Jacqmin-Gadda, H. (2016). Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: a latent process and latent class approach. *Statistics in medicine* **35,** 382–398.

Proust-Lima, C., Séne, M., Taylor, J. M., and Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research* **23,** 74–90.

Rama, R., Swaminathan, R., and Venkatesan, P. (2010). Cure models for estimating hospital-based breast cancer survival. *Asian Pacific Journal of Cancer Prevention* **11,** 387–391.

Rizopoulos, D., Molenberghs, G., and Lesaffre, E. M. (2017). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal* **59,** 1261–1276.

Rouanet, A., Joly, P., Dartigues, J.-F., Proust-Lima, C., and Jacqmin-Gadda, H. (2016). Joint latent class model for longitudinal data and interval-censored semi-competing events: Application to dementia. *Biometrics* **72,** 1123–1135.

Schemper, M. and Henderson, R. (2000). Predictive accuracy and explained variation in cox regression. *Biometrics* **56,** 249–255.

Schlattmann, P. and Böhning, D. (1993). Mixture models and disease mapping. *Statistics in Medicine* **12,** 1943–1950.

Schoop, R., Schumacher, M., and Graf, E. (2011). Measures of prediction error for survival data with longitudinal covariates. *Biometrical Journal* **53,** 275293.

Song, X., Davidian, M., and Tsiatis, A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* **58,** 742753.

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* pages 1171–1177.

Symington, F. W., Pepe, M. S., Chen, A. B., and Deliganis, A. (1990). Serum tumor necrosis factor alpha associated with acute graft-versus-host disease in humans. *Transplantation* **50,** 518–520.

Tein, J. Y., Coxe, S., and Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling* **20,** 640–657.

Tekle, F. B., Gudicha, D. W., and Vermunt, J. K. (2016). Power analysis for the bootstrap likelihood ratio test for the number of classes in latent class models. *Advances in Data Analysis and Classification* **10,** 209–224.

the Leukemia and Lymphoma Society (2016). Facts and statistics. https://www.lls.org/facts-and-statistics/facts-and-statistics-overview.

Therneau, T. and Lumley, T. (2011). Survival: Survival analysis, including penalised likelihood. r package version 2.36-5. *Survival: Survival analysis, including penalised likelihood. R package version* pages 2–36.

Therneau, T. M. (2015). *A Package for Survival Analysis in S.* version 2.38.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* pages 809–834.

Uguccioni, M., Meliconi, R., Nesci, S., Lucarelli, G., Ceska, M., Gasbarrini, G., and Facchini, A. (1993). Elevated interleukin-8 serum concentrations in beta-thalassemia and graft-versus-host disease. *Blood* **81,** 2252–2256.

van Houwelingen, H. and Putter, H. (2011). *Dynamic prediction in clinical survival analysis*. CRC Press.

van Houwelingen, H. C. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics* **34,** 70–85.

Vander Lugt, M. T., Braun, T. M., Hanash, S., Ritz, J., Ho, V. T., Antin, J. H., Zhang, Q., Wong, C.-H., Wang, H., Chin, A., et al. (2013). St2 as a marker for risk of therapy-resistant graft-versus-host disease and death. *New England Journal of Medicine* **369,** 529–539.

Vermunt, J. K. and Magidson, J. (2003). Latent class models for classification. *Computational Statistics & Data Analysis* **41,** 531–537.

Weisdorf, D., Zhang, M., Arora, M., Horowitz, M. M., Rizzo, J. D., and Eapen, M. (2012). Graft-versus-host disease induced graft-versus-leukemia effect: greater impact on relapse and disease-free survival after reduced intensity conditioning. *Biology of Blood and Marrow Transplantation* **18,** 1727–1733.

Yang, L., Yu, M., and Gao, S. (2016). Prediction of coronary artery disease risk based on multiple longitudinal biomarkers. *Statistics in Medicine* **35,** 1299–1314.

Yu, M., Law, N., Taylor, J., and Sandler, H. M. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica* **14,** 835–862.

Zhang, H. (2016). Automatic model selection for mixture models via an information theoretic approach.

Zheng, Y. and Heagerty, P. J. (2005). Partly conditional survival models for longitudinal data. *Biometrics* **61,** 379–391.