ABSTRACT

On Inferring Cognitive Impairment from a Battery of Tests and Predicting an Event-of-Interest Using Longitudinal and Time-to-Event Data

Morgan McCreary, Ph.D.

Chairperson: Dennis A. Johnston, Ph.D.

In this dissertation we first investigated two commonly used methods' and a recently proposed method's ability to predict conditional survival probabilities based on longitudinal biomarker measurements. Then, in preparation for the application of such dynamic prediction methods, we proposed the use of Bayesian multivariate mixture model accounting for censoring to infer cognitive impairment status at baseline from a battery of cognitive tests in the presence of censoring. The currently used methods for inferring cognitive impairment from a battery of cognitive tests and the proposed method were applied to the task of inferring cognitive impairment from a battery of cognitive tests administered to pediatric Multiple Sclerosis patients. The impact of censoring on the inferred cognitive impairment status was examined, as well as the predictive accuracy of the proposed and currently used methods via simulation. Finally, in order to infer cognitive impairment based on results from a battery of cognitive tests obtained during follow-up in the presence of a practice effect, we proposed the use of a Bayesian continuous-time mixed hidden Markov model. To account for the practice effect in the hidden Markov model, we proposed incorporating an adapted form of the half-life regression equation presented by Settles and Meeder. We simulated two example datasets based on the results from the analysis of the

battery of cognitive tests at baseline due to the lack of longitudinal cognitive testing data from pediatric Multiple Sclerosis patients and applied the hidden Markov model to infer cognitive impairment. We examined the ability of the hidden Markov model to correctly infer cognitive impairment at each time point during follow-up for the simulated patients as well as the accuracy of parameter estimates. The predictive accuracy of the proposed methods in simulated and real data obtained from pediatric Multiple Sclerosis patients enable us to efficiently design clinical studies and trials aimed at improving the understanding and treatment of cognitive impairment in this patient population, as well as other diseases in which cognitive impairment is a known effect.

On Inferring Cognitive Impairment from a Battery of Tests and Predicting an Event-of-Interest Using Longitudinal and Time-to-Event Data

by

Morgan McCreary, B.S., M.S.

A Dissertation

Approved by the Department of Statistical Science

_____

Jack D. Tubbs, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

_____

Dennis A. Johnston, Ph.D., Chairperson

_____

James D. Stamey, Ph.D.

_____

Dean M. Young, Ph.D.

_____

F. Carson Mencken, Ph.D.

Accepted by the Graduate School
August 2018

_____

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Thank you Dr. Johnston for your continued assistance and patience along this process and all the effort you put forth to help me arrive at this point. Thank you committee members for your willingness to review the culmination of our work and offer any welcomed suggestion. Finally, thank you Dr. Benjamin Greenberg; without your guidance and support I would not be where I am today.

DEDICATION

To my parents, family, and friends; without your continued support and patience none of this would have been possible.

CHAPTER ONE

Introduction

Our initial intent for this dissertation was to examine methods for the dynamic prediction of longitudinal and time-to-event data in which real-time predictions are made of the probability a patient will experience an event-of-interest based on longitudinal biomarker measurements (Van Houwelingen and Putter, 2011; Rizopoulos, 2012; Huang et al., 2016). As the medical field continues its transition to individualized medicine, dynamic prediction has become a valuable tool given its ability to predict a given patient's diseases prognosis based on biomarker measurements and influence future care. Recently, Huang et al. (2016) proposed a novel method for dynamic prediction of longitudinal and time-to-event data which addressed shortcomings of the two commonly used methods, landmark analysis and joint modeling of longitudinal and time-to-event data. The simulation study constructed by Huang et al. (2016) examined the ability of the three models to predict the conditional survival probabilities a given patient experiences the event-of-interest before some future time point given that they have not experienced the event-of-interest based on longitudinal biomarker measurements. However, their simulation trained and tested on the same group of patients, potentially resulting in overestimated predictive accuracy as a consequence of overfitting. Additionally, Huang et al. (2016) did not examine the predictive accuracy of the three methods for differing lengths of time between the time up to which the longitudinal measurements are obtained and the future time point. Also, the biologic variability of the simulated longitudinal biomarker in Huang et al. (2016) was relatively small compared to the simulated biomarker values. Therefore, in Chapter Two we reconstructed the simulation study from Huang et al. (2016) and examined the predictive accuracy of the model trained on a training sample to predict

the conditional survival probability of a testing sample of patients. Additionally, we examined the predictive accuracy for various lengths of time between the time up to which the longitudinal measurements were obtained and the future time point along with differing degrees of biologic variability.

Following examination of the three dynamic prediction methods, our interest was in developing an improved method to infer the cognitive impairment from a battery of cognitive tests in preparation of the application of dynamic prediction methods to this type of data as well the analysis of datasets currently being collected. The current methods used to infer cognitive impairment violate many of the underlying standard assumptions. Cognitive impairment is a latent variable inferred based on a given patient's scores from a battery, or collection, of cognitive tests which are often correlated. However, the statistical methods currently used to infer cognitive impairment operate under the assumption of independence amongst the cognitive tests in a given battery. Additionally, censored cognitive test scores are commonly encountered in clinical data and can lead to biased parameter estimates and incorrect inferred cognitive status. Therefore, we proposed the use of a Bayesian multivariate finite mixture model accounting for censoring to estimate the correlation amongst the test and infer each patient's cognitive impairment status in the presence of censored test scores. In Chapter Three we inferred patients' cognitive impairment status from a battery of cognitive tests administered to pediatric Multiple Sclerosis patients at the Pediatric Demyelinating Diseases Clinic of Children's Health Dallas using the currently used methods and the Bayesian multivariate finite mixture model in the presence and absence of censoring. Additionally, we have constructed a simulation study to examine the sensitivity and specificity of the currently used methods and the Bayesian multivariate finite mixture model to infer cognitive impairment, as well as agreement in the inferred cognitive status amongst the various methods.

While the Bayesian multivariate finite mixture model accurately inferred cogni-

tive impairment at baseline in the simulation study, we needed to extend the Bayesian multivariate finite mixture model to infer cognitive impairment based on scores from a battery of cognitive tests during follow-up testing. This task was further complicated by the well-documented presence of a practice effect for many cognitive tests in which a patient's score increases due to repeated exposure to a given cognitive test (Dikmen et al., 1999; Beglinger et al., 2005; Collie et al., 2003; Calamia et al., 2012). The practice effect can create the artificial perception that a patient's cognitive test score increased. Therefore, we needed to account for the practice effect to accurately infer a given patient's cognitive impairment status during follow-up. To infer a patient's cognitive impairment status during follow-up, we have proposed the use of a Bayesian continuous-time mixed hidden Markov model. To account for the practice effect during repeated testing, an additional term adapted from the half-life regression equation proposed by Settles and Meeder (2016) was incorporated into the Bayesian continuous-time mixed hidden Markov model. Due to the lack of longitudinal scores for the battery of cognitive test of interest from pediatric Multiple Sclerosis patients, data were simulated based on results presented in Chapter Three and the parameter estimates corresponding to each subpopulation and the predictive accuracy of the inferred cognitive impairment status were examined in Chapter Four.

CHAPTER TWO

Current Methods for Dynamic Prediction of Longitudinal and Time-to-Event Data

2.1 Introduction

Dynamic prediction is an attempt to make real-time predictions of a given patient's disease prognosis based on longitudinal biomarker measurements (Huang et al., 2016). Dynamic prediction has become increasingly used in the prediction of various cancers, including Chronic Myeloid Leukemia (CML), Advanced Ovarian Cancer, and Breast Cancer (Huang et al., 2016; Van Houwelingen and Putter, 2011; Rizopoulos, 2012). Estimation of the conditional survival probability based on longitudinal biomarker measurements, whose values are thought to be a manifestation of a patient's given disease state, enables physician to make real-time clinical decisions to limit further disease progression.

Two common methods for dynamic prediction of longitudinal and time-to-event data are landmark analysis and joint modeling of longitudinal and time-to-event data (Van Houwelingen and Putter, 2011; Rizopoulos, 2012). However, associated with each of these methods are specific shortcomings that limit their predictive accuracy. Recently, Huang et al. (2016) proposed a novel method for the estimation of the conditional survival based on longitudinal measurements to address the shortcomings of the two aforementioned techniques. A simulation study was constructed in Huang et al. (2016) to examine the predictive accuracy of the two commonly used methods and their proposed method. Unfortunately, the design of the simulation study was flawed and the results may provide inaccurate estimates of the predictive accuracy of three methods. Also, the authors only considered a relatively low level of biologic variability in their simulated longitudinal biomarker measurements. Lastly, the authors did not examine the predictive accuracy of the three methods for different

lengths of time between the time up to which the longitudinal measurements were obtained and the future time point. Therefore, we have addressed the flaws of the simulation study presented in Huang et al. (2016) and have presented those results here.

Section 2.2 provides a detailed description of the basic survival notation that will be used in the remainder of this chapter. Section 2.3 provides information regarding the two commonly used dynamic prediction methods and the method recently proposed by Huang et al. (2016). Section 2.4 then describes the simulation study presented in Huang et al. (2016) and the changes we have made to obtain more correct and informative results, as well as the results obtained in our simulation study. Finally, Section 2.5 describes the apparent shortcomings of the three methods investigated here.

2.2 Basic Survival Notation

Suppose n patients with a particular disease are enrolled in a clinical study and followed for a previously set period of time. Let T_i denote the true time of an event-of-interest occurring after the initiation of follow-up for the ith patient, $i = 1, \ldots, n$. Events-of-interest in clinical studies often are defined as disease relapse or death due to disease. However, during the course of follow-up the particular event-of-interests may not be observed in a subgroup of patients who do not experience the event-of-interest during their study enrollment. The time to the event-of-interest for these subjects is said to be censored. More specifically, patients who are lost to follow-up during the course of a study are said to be right censored and patients who complete the study event-free are said to be administratively censored. In either case, T_i is not observed for these patients, only the censoring time for the ith patient, C_i, is observed. Therefore, we will define the random variable $X_i = \min(T_i, C_i)$ and the censoring indicator $\Delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ is the indicator function. We will denote a

realization of the random variable pair $\{X_i, \Delta_i\}$ as $\{x_i, \delta_i\}$. A crucial assumption for the validity of the model estimates we will discuss is that of independent censoring. Under independent censoring, the true survival time T_i and the censoring time C_i are independent. Methods for dependent censoring exists and are discussed in Collett (2015), but are not discussed further here.

The true time to an event-of-interest for the ith patient, T_i, is a non-negative random variable with probability density function $f(t)$. We will then define cumulative incidence function evaluated at t to be the cumulative probability of an event-of-interest occurring before t, calculated as

$$F(t) = P(T < t) = \int_0^t f(u)du. \tag{2.1}$$

From (2.1), we arrive at the survivor function,

$$S(t) = P(T \geq t) = 1 - F(t),$$

which is the probability that the event-of-interests occurs at or after time t. An additional function in which we will be interested is the hazard function, $h(t)$. The hazard function for the ith patient represents the instantaneous risk of the event-of-interest occurring at time t, given that it has not occurred before time t, and is defined as

$$h_i(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T_i < t + \delta t | T_i \geq t)}{\delta t} \right\}.$$

Similarly, the cumulative hazard function for the ith patient is defined as

$$H_i(t) = \int_0^t h_i(u)du, \tag{2.2}$$

and represents the cumulative risk of the event-of-interest occurring by time t, given that the event-of-interest has not occurred prior to t. It can be shown that

$$h_i(t) = \frac{f_i(t)}{S_i(t)}, \tag{2.3}$$

and it follows from (2.3) and (2.2) that $h_i(t) = -\frac{d}{dt}\{\log S_i(t)\}$, $S_i(t) = \exp\{-H_i(t)\}$, and $H_i(t) = -\log S_i(t)$.

An additional measure related to the survivor function that is often of interest is the conditional probability of survival up to a time point $t+u$ given survival up to time t, $u > 0$. This probability can be expressed as

$$\begin{aligned}\pi(t+u|t) &= P(T_i \geq t+u | T_i \geq t, u > 0) \\ &= \frac{P(T_i \geq t+u \cap T_i \geq t | u > 0)}{P(T_i \geq t)} \\ &= \frac{P(T_i \geq t+u | u > 0)}{P(T_i \geq t)} \\ &= \frac{S_i(t+u)}{S_i(t)}. \end{aligned} \tag{2.4}$$

This quantity will be our main quantity-of-interest as we delve into dynamic prediction in later sections.

Suppose that in addition to recording time to the event-of-interest, the clinical study also collects additional covariates thought to be related to T_i. We will be interested in covariates of two types: time-independent covariates and endogenous time-dependent covariates. Time-independent covariates include baseline information that does not change during the course of the clinical study such as gender, age at onset, and treatment administered throughout the course of the clinical study. We will denote the p-dimensional vector of time-independent covariates for the ith patient $Y_i = (Y_{i1}, \ldots, Y_{ip})^\top$. Endogenous, or internal, time-dependent covariates are time-dependent measurements obtained for each subject that relate to T_i and can only be measured while the patient is alive. An example of endogenous time-dependent covariates are longitudinal biomarker measurements obtained from a patient during the course of the clinical study. We will denote the q-dimensional vector of endogenous time-dependent covariates for the ith patient at time t as $Z_i(t) = (Z_{i1}(t), \ldots, Z_{iq}(t))^\top$.

In order to model time-to-event data with endogenous time-dependent covariates and, thereby, estimate the survivor function, the Cox regression model with

endogenous time-dependent covariates can be used (Cox, 1992, 1975). Using the notation presented above, the Cox regression model containing time-independent and endogenous time-dependent covariates is defined as

$$h_i(t) = h_0(t) \exp\left\{\boldsymbol{\alpha}^\top Y_i + \boldsymbol{\gamma}^\top Z_i(t)\right\},$$

such that $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$ is a p-dimensional vector of unknown coefficients corresponding to the time-independent covariates, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^\top$ is a q-dimensional vector of unknown coefficients corresponding to the endogenous time-dependent covariates, and $h_0(t)$ is the baseline hazard function, defined as the value of the hazard function when all covariates in the model are zero. The resulting survivor function for the ith individual is then defined as

$$\begin{aligned} S_i(t) &= \exp\{-H_i(t)\} \\ &= \exp\left\{-\int_0^t h_i(u)du\right\} \\ &= \exp\left\{-\int_0^t h_0(u) \exp\left\{\boldsymbol{\alpha}^\top Y_i + \boldsymbol{\gamma}^\top Z_i(u)\right\} du\right\}. \end{aligned} \tag{2.5}$$

Estimation of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ and the baseline hazard function, $h_0(t)$, are complicated by the presence of endogenous time-dependent covariates. Additionally, estimation of the survivor function in equation (2.5) is complicated due to the presence of the integral from 0 to t. Particularly, this presents a problem for endogenous time-dependent covariates such as longitudinal biomarker measurements, whose value is often not known at all time points $u \in [0, t]$. Estimation of $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, $h_0(t)$, and $S_i(t)$ remains a topic of interest in the analysis of time-to-event and longitudinal data research, with a wide array of methods differing in their approach. We will forego an explanation of such methods until a later section.

2.3 Dynamic Prediction of Time-to-Event and Longitudinal Data

The main quantity of interest in dynamic prediction is the conditional probability that a patient survives up to a time $t + u$, $u > 0$, given that a patient

has survived up to time t with longitudinal biomarker measurements available at times $t_{i1}, \ldots, t_{ik} \in [0, t]$ and known time-independent covariates. We will denote the $q \times k$-dimensional matrix of biomarker measurements for the ith subject as $\mathbf{Z}_i = (Z_{ij}(t_{i1}), \ldots, Z_{ij}(t_{ik}))$, $j = 1, \ldots, q$. Using equation (2.4), we can express the quantity of interest in dynamic prediction as

$$\pi(t + u|t) = \frac{S_i(t + u|u > 0, \mathbf{Z}_i, Y_i)}{S_i(t|\mathbf{Z}_i, Y_i)}.$$

Therefore, estimation of the quantity of interest requires estimation of the survivor distributions at time $t + u$ and t, given both time-independent and endogenous time-dependent covariates. As discussed previously, the presence of endogenous time-dependent covariates creates difficulties in the estimation of the survivor function. A brief overview of the landmark approach, joint modeling of longitudinal and time-to-event approach, and the Huang et al. (2016) two-stage approach are provided below. For the interested reader, further information regarding landmark analysis, joint modeling of longitudinal and time-to-event data, and the Huang et al. (2016) two-stage approach can be found in Van Houwelingen and Putter (2011), Rizopoulos (2012), and Huang et al. (2016), respectively.

2.3.1 Landmark Analysis

As presented in Van Houwelingen and Putter (2011) and Van Houwelingen (2007), the landmark analysis approach uses the covariate information at the selected landmark t for only those patients who have yet to experience the event-of-interest or were lost to follow-up before t. The hazard function for the ith patient at time $t + u$ based on covariates at landmark time t is defined as

$$h_i(t + u|Y_i, Z_i(t)) = h_{t,0}(t + u) \exp\{\theta_{LM}(t)Y_i + \gamma_{LM}(t)Z_i(t)\}, \quad u > 0, \tag{2.6}$$

where $\theta_{LM}(t)$ and $\gamma_{LM}(t)$ are not time-varying coefficients, but instead the coefficients corresponding to the time-independent and time-dependent covariate at the

landmark time t. Additionally, $h_{t,0}(t+u)$ is the baseline hazard function at time $t+u$ based on the covariates at time t. To estimate $P(T \geq t+u|T \geq t, Y_i, Z_i(t))$, the landmark approach first estimates the parameters $\theta_{LM}(t)$ and $\gamma_{LM}(t)$ in equation (2.6) by maximizing the partial likelihood of the the Cox regression model, treating $Z_i(t)$ as a time-independent covariate, defined as

$$L(\theta_{LM}(\theta), \gamma_{LM}(t)) = \prod_{i=1}^{n} \left\{ \frac{\exp\{\theta_{LM}(t) y_i + \gamma_{LM}(t) z_i(t)\}}{\sum_{\ell \in R(x_i)} \exp\{\theta_{LM}(t) y_\ell + \gamma_{LM}(t) z_\ell(t)\}} \right\}^{\delta_i},$$

where $R(x_i)$ is the risk set defined as the set of subjects who have not experienced the event-of-interest prior to x_i and have not been censored prior to x_i (Collett, 2015). Then, the baseline survival function, or survival function when all are zero, is estimated using the Breslow estimate defined as

$$\widehat{S}_0(t) = \exp\left\{ - \sum_{i:x_i \leq t} \left[\frac{\delta_i}{\sum_{j:x_j \geq x_i} \exp(\widehat{\theta}_{LM}(t) y_j + \widehat{\gamma}_{LM}(t) z_j(t))} \right] \right\},$$

and the estimated baseline hazard function at time $t+u$ in the landmark approach is defined as

$$\widehat{S}_0(t+u) = \exp\left\{ - \sum_{i:x_i \leq t+u} \left[\frac{\delta_i}{\sum_{j:x_j \geq x_i} \exp(\widehat{\theta}_{LM}(t) y_j + \widehat{\gamma}_{LM}(t) z_j(t))} \right] \right\}.$$

Then, for a new patient with time-independent covariate values y_{new} and time-dependent covariate values $z_{\text{new}}(t)$, the conditional probability of not experiencing the event-of-interest up to time $t+u$, given that the new patient has not experienced the event-of-interest before time t, is estimated using

$$\widehat{\pi}(t+u|t) = \left[\frac{\widehat{S}_0(t+u)}{\widehat{S}_0(t)} \right]^{\exp(\widehat{\theta}_{LM}(t) y_{\text{new}} + \widehat{\gamma}_{LM}(t) z_{\text{new}}(t))}.$$

The landmark analysis approach is the simplest of the three methods considered here, a very appealing feature of this approach. Van Houwelingen (2007) argued that the landmarking approach can achieve comparable results without the need to create complex models. However, the landmark approach requires for all patients

to have measurements collected at the selected landmark t, a requirement that is often unreasonable in the clinical setting. Imputation techniques can be used to obtain estimates of covariates at the selected landmark time t, but can lead to biased estimates (Tsiatis and Davidian, 2001).

2.3.2 Joint Modeling of Longitudinal and Time-to-Event Data

The joint modeling of longitudinal and time-to-event data approach attempts to address the issue previously mentioned in the estimation of $S_i(t)$ in equation (2.5) in that biomarker measurement are rarely available for all time points from baseline to t, and unavailable after time t to $t+u$, $u>0$. The proposed solution offered by the joint modeling approach is to estimate a model for the biomarker measurement over time based on the observed biomarker measurements, where for the ith patient it is assumed that

$$Z_i(t) = f_i(t) + \epsilon_i(t), \quad i = 1, \ldots, n,$$

where $f_i(t)$ represents the underlying true distribution of the longitudinal biomarker measurements for the ith patient and $\epsilon_i(t)$ is associated with any measurement and/or biological variability associated with the biomarker measurements. The function $f_i(t)$ is often modeled using linear mixed effects models such that

$$\begin{aligned} f_i(t) &= u_i^\top(t)\beta + v_i^\top b_i, \\ b_i &\sim \mathcal{N}(0, D), \\ \epsilon_i(t) &\sim \mathcal{N}(0, \sigma^2), \end{aligned}$$

where $u_i(t)$ is the design vector for the fixed effects β and $v_i(t)$ is the design vector for the random effects b_i. The joint modeling hazard function and survival function for the ith patient at time t is then defined as

$$h_i(t|Y_i, f_i(t)) = h_0(t)\exp\{\theta^\top Y_i + \gamma^\top f_i(t)\}, \tag{2.7}$$

and

$$S_i(t|Y_i, f_i(t)) = \exp\left(-\int_0^t h_0(s)\exp\{\theta^\top Y_i + \gamma^\top f_i(s)\}ds\right). \tag{2.8}$$

Additionally, in joint modeling, the baseline hazard function, $h_0(t)$, in equations (2.7) and (2.8) must be specified. Rizopoulos (2012) discusses various options for specifying the baseline hazard function and mentions a simple option in the piecewise-constant mode. In the piecewise-constant model, the baseline hazard function is defined as

$$h_0(t) = \sum_{q=1}^{Q} \xi_q I(\nu_{q-1} < t < \nu_q), \tag{2.9}$$

such that $I(\cdot)$ denotes the identity function, $0 = \nu_0 < \nu_1 < \cdots < \nu_Q$ denotes points along the time interval of interest, with ν_Q being greater than largest observed time or the end of the study, and ξ_q denoting the value of the hazard within the interval $(\nu_{q-1}, \nu_q]$ (Rizopoulos, 2012).

The traditional method for estimation of the joint model is based on maximization of the likelihood function of the joint distribution of $(X_i, \Delta_i, Y_i, Z_i(t))$ using the Expectation-Maximization (EM) algorithm or the Newton-Rhapson algorithm (Rizopoulos, 2012). We do not provide the likelihood (or log-likelihood) function or EM algorithm for maximization of the likelihood (or log-likelihood). Instead we direct the reader to Rizopoulos (2012, Chapter 4), which provides the log-likelihood function corresponding to the joint distribution of $(X_i, \Delta_i, Y_i, Z_i(t))$ as well as the steps of the EM algorithm used to maximize the log-likelihood function.

Rizopoulos (2012) has proposed the use of Monte Carlo simulation schemes to estimate the conditional probability $\pi(t + u|t)$. Again, we will not reproduce the Monte Carlo simulation scheme here but instead direct the reader to Rizopoulos (2012, Chapter 7) for a detailed description of the Monte Carlo simulation.

2.3.3 Huang et al. (2016) Two-Stage Approach

Huang et al. (2016) proposed a two-stage approach for dynamic prediction of longitudinal and time-to-event data to address shortcomings of the landmark approach and the joint modeling approach. Specifically, the authors wanted a method which did not require subjects to have biomarker measurements obtained at the same time points, as is required by the landmark approach, and did not require the specification of the longitudinal biomarker trajectory model, as does the joint modeling approach, due to the potential impact of misspecification.

The Huang et al. (2016) two-stage approach assumed a Cox regression model for the hazard function for the ith patient, defined as

$$h_i(t|Y_i, Z_i(t)) = h_0(t) \exp\{\theta^\top Y_i\} \exp\{\gamma^\top(t) Z_i(t)\},$$

where $\gamma(t)$ is a time-varying coefficient. Additionally, the authors specified the covariate Y_i to include the baseline value of the biomarker measurements, as well as time-independent covariates. Therefore, the model proposed by Huang et al. (2016) does not require specification of the model of the longitudinal biomarker trajectory, but does require the specification of the model of the time-varying coefficient. The first stage of the two-stage approach estimates $h_0(t)$ and θ, ignoring the time-dependent biomarker measurements. In order to first estimate θ, the partial likelihood, defined as

$$L(\theta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\theta^\top Y_i)}{\sum_{j:x_j \geq x_i} \exp(\theta^\top Y_j)} \right\},$$

is maximized. Then, using the estimated value for θ, $\widehat{\theta}$, the baseline hazard function and baseline survival function are estimated using the Breslow estimator; that is,

$$\widehat{h}_0(x_i) = \frac{\delta_i}{\sum_{j:x_j \geq x_i} \exp(\widehat{\theta}^\top y_i)}, \tag{2.10}$$

and

$$\widehat{S}_0(t) = \exp\left\{-\sum_{i:x_i \leq t} \widehat{h}_0(x_i)\right\}.$$

The goal of the second stage of the two-stage approach is to estimate the time-varying coefficient associated with the longitudinal biomarkers, $\gamma(t)$. The authors specified a model for $\gamma^\top(t)Z_i(t)$ defined as

$$\gamma^\top(t)Z_i(t) = \gamma_0^\top(t) + \gamma_1^\top Z_i(t),$$

where $\gamma_0(t)$ and $\gamma_1(t)$ are defined as fractional polynomials such that

$$\gamma_k(t) = \gamma_{k0} + \gamma_{k1}\ln(t) + \gamma_{k2}\sqrt{t} + \gamma_{k3}\frac{1}{\sqrt{t}} + \gamma_{k4}t + \gamma_{k5}\frac{1}{t} + \gamma_{k6}t^2 + \gamma_{k7}\frac{1}{t^2},$$

for $k = 1, 2$. The authors then removed higher order polynomials using backwards elimination.

Huang et al. (2016) then defined their likelihood function for the ith patient at time t as

$$\widehat{L}_{ij}(\gamma(t)) = [\widehat{h}_0(x_i)\exp\{\widehat{\theta}^\top y_i + \gamma^\top z_i(t)\}]^{\delta_i} \times \exp\left[-\exp\{\widehat{\theta}^\top y_i + \gamma^\top(t)z_i(t)\}\sum_{m:t<x_m\leq x_i} h_0(x_m)\right].$$

Then, under the working independence assumption among the different time points t_{ij}, the 'working' log-likelihood function is defined as

$$\widehat{\ell}(\gamma(t)) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\delta_i\{\log[\widehat{h}_0(x_i)] + \widehat{\theta}^\top y_i + \gamma^\top(t_{ij})z_i(t_{ij})\} - \exp\{\widehat{\theta}^\top y_i + \widehat{\gamma}^\top(t_{ij})z_i(t_{ij})\}\sum_{m:t_{ij}<x_m\leq x_i}\widehat{h}_0(x_m). \quad (2.11)$$

However, given that $\widehat{h}_0(x_i)$ is calculated according to equation (2.10), if $\delta_i = 0$, then $\widehat{h}_0(x_i) = 0$ and $\delta_i \log[\widehat{h}_0(x_i)]$ in equation (2.11) is indeterminate. Therefore, we will use a corrected version of the log-likelihood which does not encounter this issue.

Instead of using the 'working' log-likelihood function defined in equation (2.11), we define the 'working' log-likelihood to be

$$\widehat{\ell}(\gamma(t)) = \sum_{i=1}^{r}\sum_{j=1}^{n_i} \log[\widehat{h}_0(x_i)] + \widehat{\theta}^\top y_i + \gamma^\top(t_{ij}) z_i(t_{ij}) \\ - \sum_{i=1}^{n}\sum_{j=1}^{n_i} \exp\{\widehat{\theta}^\top y_i + \widehat{\gamma}^\top(t_{ij}) z_i(t_{ij})\} \sum_{m: t_{ij} < x_m \leq x_i} \widehat{h}_0(x_m), \quad (2.12)$$

where r denotes those subjects who are not censored during the course of the study. Then, the maximum likelihood estimate of $\gamma(t)$ is obtained by maximizing the log-likelihood in equation (2.12).

Finally, for a new patient who has time-independent measurements y_{new} and biomarker values at time t $z_{\text{new}}(t)$, the conditional survival probability is estimated to be

$$\widehat{\pi}(t+u|t) = \left[\frac{\widehat{S}_0(t+u)}{\widehat{S}_0(t)}\right]^{\exp\{\widehat{\theta}^\top y_{\text{new}} + \widehat{\gamma}^\top(t) z_{\text{new}}(t)\}}.$$

2.4 Simulation

2.4.1 Generation of Longitudinal and Time-to-Event Data

In order to examine the predictive accuracy of the landmark approach, joint modeling approach, and the Huang et al. (2016) two-stage approach for varying degrees of biological variability of the biomarker, we have constructed a simulation study consisting of $n = 200$ patients followed for a period of 10 years with biomarker measurements taken at baseline and every year over the study duration. To generate the longitudinal and time-to-event data, we first generate the longitudinal biomarker measurements from the following linear mixed effects model

$$f_i(t_{ij}) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) t_{ij}$$
$$Z_i(t_{ij}) = f_i(t_{ij}) + \epsilon_i(t_{ij}),$$

where $\beta_0 = 3$ and $\beta_1 = 2$. The vector of random effects, $\boldsymbol{b}_i = (b_{0i}, b_{1i})^\top$, were given a multivariate normal distribution with mean vector $(0, 0)^\top$ and covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{b}} = \begin{bmatrix} 4 & 0.1 \\ 0.1 & 2 \end{bmatrix}.$$

Lastly, for the linear mixed effects model, the error term was assumed to be attributed to biologic variation in the biomarker, as opposed to measurement error. Additionally, for the ith patient the error term at each time-point is assumed to be independent and distributed $\mathcal{N}(0, \sigma)$. In the original simulation study constructed by Huang et al. (2016) to investigate these three methods of dynamic prediction, the authors considered a low level of biologic variability, relative to the values of $Z_i(t)$ based on the values of β_0 and β_1, by assuming $\sigma = 0.6$. Therefore, in addition to investigating the predictive accuracy of the three model for longitudinal data with a biologic variability of $\sigma = 0.6$, we will also investigate greater degrees of biologic variability with $\sigma = 2$ and $\sigma = 4$. Simulated longitudinal biomarker trajectories for 8 random patients for each value of σ are shown in Figure 2.1.

Huang et al. (2016) assumed that the hazard function for the ith patient followed the Cox regression model

$$h_i(t|Y_i, Z_i(t)) = h_0(t) \exp\{\psi f_i(t)\}, \tag{2.13}$$

where ψ was defined as a method to control how the survival time was influenced by the longitudinal biomarker values and were set at two different values of $\psi = 0.8$ and $\psi = 1.6$. While increasing the value of ψ increased the root-mean-squared prediction error of the estimated conditional survival, it did not affect the performance of each method relative to the others. That is, if the ordering of root-mean-squared prediction error from smallest to largest were the joint modeling approach, Huang et al. (2016) two-stage approach, and the landmark approach, this ordering did not change for an increase in ψ. Therefore, we chose to exclude the ψ term and assume that the hazard

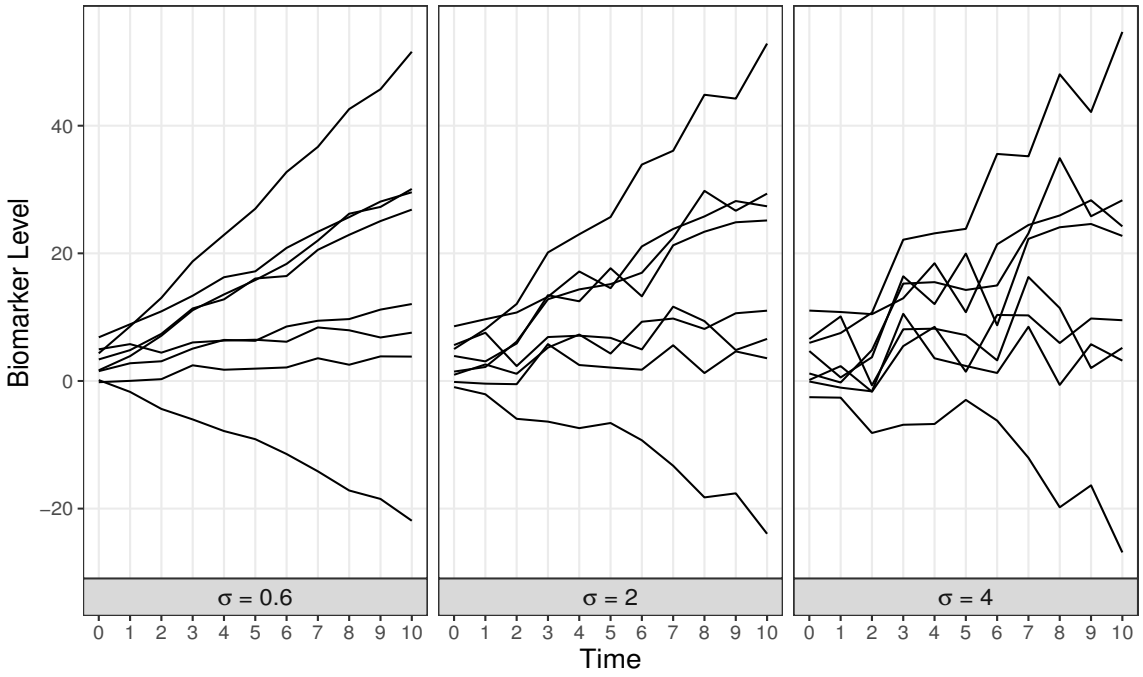


Figure 2.1. Simulated biomarker trajectories from 8 patients with biologic variability equal to $\sigma = 0.6$, $\sigma = 2$, and $\sigma = 4$.

function for the ith patient follows the Cox regression model

$$h_i(t|Y_i, Z_i(t)) = h_0(t) \exp\{f_i(t)\},$$

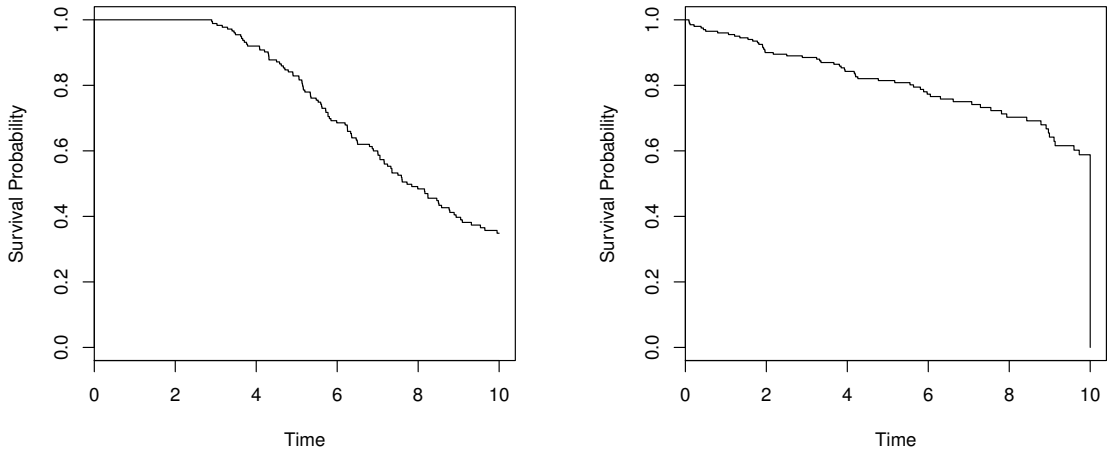
where $h_0(t) = \lambda\nu e^{\nu-1}$, a Weibull hazard function. The shape parameter of the Weibull hazard function, ν, was set to be 0.5 and the scale parameter, λ, was to be $\lambda = 5\times10^{-6}$ in order to achieve a 50% censoring rate.

Recall that cumulative incidence function can be represented as

$$\begin{aligned} F_i(t) &= 1 - S_i(t) \\ &= 1 - \exp\{-H_i(t)\}. \end{aligned}$$

If we then assume that $F(t) \sim Uniform(0, 1)$, we can use the inverse transform method to obtain the failure time for the ith patient based on the following:

$$P(F(T_i) < u) = P(T_i < F^{-1}(u)) = u,$$



(a) (b)

Figure 2.2. Kaplan-Meier curves corresponding to (a) the event-of-interest and (b) censoring.

for $u \in [0, 1]$. Hence, if $U \sim Uniform(0, 1)$, then $F^{-1}(U)$ has the same distribution as T_i. As in Beyersmann et al. (2011), we conducted the inverse transform method by first computing $F^{-1}(U) = H^{-1}(-\ln(1-u))$. However, the inverse of H does not have a closed form expression. Therefore, we are required to compute it numerical inversion. Then, we can generate the random variable for U such that $U \sim$ Uniform$(0, 1)$. It then follows that for the ith patient, $H^{-1}(-\ln(1-u_i)) := t_i$. Then, the censoring variable for the ith patient, C_i, will be generated from the Uniform$(0, 28)$ distribution and X_i will then be defined to be $\min(C_i, X_i)$. Finally, $\Delta_i = I(T_i < C_i)$. The resulting Kaplan-Meier curve for the event-of-interest and censoring for data simulated for all three values of σ are shown in Figure 2.2.

Table 2.1 provides the number of simulated patients who failed or were censored during each interval.

2.4.2 Results

In the original article by Huang et al. (2016), the authors do not describe at which time point t biomarker measurements were used to predict $\pi(T_i \geq 10|T_i > t)$.

Table 2.1. Number of simulated patient who failed, were censored, or remained in the risk set during each time interval.

Time Interval	Failure	Censored	Remaining
$[0, 1)$	0	8	192
$[1, 2)$	0	12	180
$[2, 3)$	2	3	175
$[3, 4)$	12	8	155
$[4, 5)$	15	5	135
$[5, 6)$	23	6	106
$[6, 7)$	13	3	90
$[7, 8)$	17	5	68
$[8, 9)$	12	5	51
$[9, 10)$	6	4	41

Additionally, the authors do not specify whether a testing and training sample were used, so we assume that the authors estimated the parameters and corresponding hazard function using the full dataset and then estimated conditional survival of all patients based on these estimates. Instead, we will examine the performance of these three methods using a testing and training sample in each iteration of the simulation.

Based on the $n = 200$ simulated patients, the three methods were used to predict $\pi(T_i \geq 10 | T_i > t)$, for $t = 1, 2, \ldots, 9$, on a testing sample of 10% of the patients remaining at the risk set at each time t. Additionally, the proportion of patients in the testing sample who were administratively censored (and experienced the event-of-interest) were designed to be equivalent to the proportion of patients in each simulated dataset who were administratively censored (and experienced the event-of-interest) after time t. For instance, at $t = 5$, 135 patients remain in the risk set, of which 94 patients experience the event-of-interest or are lost to follow-up before $t = 10$ and 41 complete the study without the event of interest. In this instance, the testing sample would contain 9 patients who experienced the event-of-interest and 4 who were administratively censored. Given that we know the true

biomarker trajectory, $f_i(t)$, for each patient at all times $t \in [0, 10]$, we can compute the true value of $\pi(T_i \geq 10|T_i > t)$ for $t = 1, 2, \ldots, 9$ based on equation (2.5). Then, to test the predictive accuracy of the landmark approach, the joint modeling approach, and the two-stage approach proposed by Huang et al. (2016), we can compute root-mean-squared prediction error (RMSE) for a testing sample of patients remaining in the risk-set at time t based on the parameter estimates obtained using a training sample of patients as follows

$$RMSE = \sqrt{[\widehat{\pi}(T_i \geq 10|T_i > t) - \pi(T_i \geq 10|T_i > t)]^2}, \qquad t = 1, 2, \ldots, 9,$$

where $\widehat{\pi}(T_i \geq 10|T_i > t)$ is the predicted conditional survival probability obtained using the landmark approach, joint modeling approach, or two-stage approach proposed by Huang et al. (2016).

The landmark approach, joint modeling approach, and two-stage approach proposed by Huang et al. (2016) were performed as discussed in Section 3.4. For the joint modeling approach, the procedure was carried out using the JM package in R (Rizopoulos, 2010). Additionally, in the joint modeling approach, the baseline hazard function $h_0(t)$ was given the piecewise-constant model shown in equation (2.9) with the number of knots specified such that the number of linear predictors in the mixed effects model and the number of knots is less than 1/10 the total number of events in the sample, as suggested by Rizopoulos (2012) to avoid overfitting. Given that the main interest in our simulation study is to investigate the effect of increasing biologic variability and not the bias introduced by misspecification of the linear mixed effects model, the linear mixed effects model in the joint modeling approach was specified to be $f_i(t) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t$ to match the true biomarker trajectory.

The RMSE for each of the three dynamic prediction methods from the simulation study for $\sigma = 0.6$, 2, and 4 are shown in Figures 2.3 - 2.5. Additionally the average RMSE and standard deviations can be found in Tables 2.2 - 2.4. As would

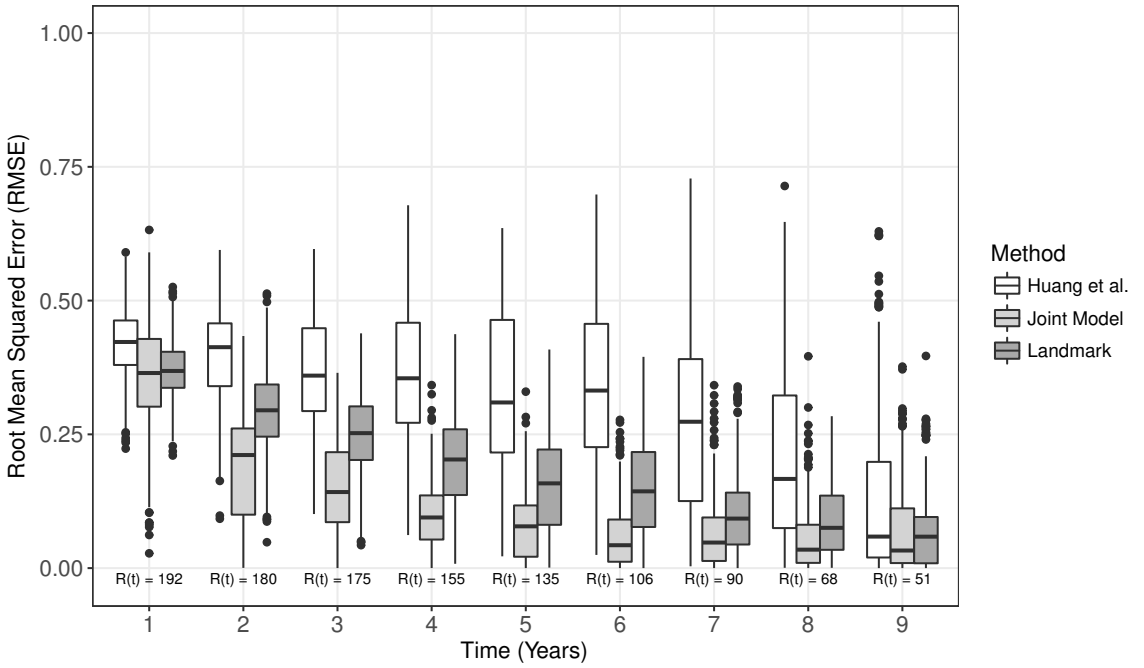


Figure 2.3. Root-mean-squared prediction error of $\pi(t + u = 10|t)$ for the landmark approach, joint modeling approach, and Huang et al. (2016) two-stage approach for times $t = 1, 2, \ldots, 9$ when $\sigma = 0.6$

be expected, the RMSE for the joint modeling approach, landmark approach, and two-stage approach proposed by Huang et al. (2016) decreases as t gets closer to $t + u = 10$. Also as would be expected, we see that the RMSE for each of the three methods increase as the value of σ increases.

When we compare the performance of the three methods over time when the biologic variability is low ($\sigma = 0.6$), we see that the joint modeling approach results in the lowest RMSE among the three methods, with the average RMSE of the Huang et al. (2016) two-stage approach having greater than a two standard deviation difference for times $t = 2$, $t = 3$, and $t = 4$ relative to the joint modeling approach while the average RMSE corresponding to the landmark approach is within two standard deviations of the average RMSE of the joint modeling approach for all times $t \in \{1, 2, \ldots, 9\}$. However, we did not find that for any $t \in \{1, 2, \ldots, 9\}$ the Huang et al. (2016) two-stage performance had greater than a two-standard deviation dif-

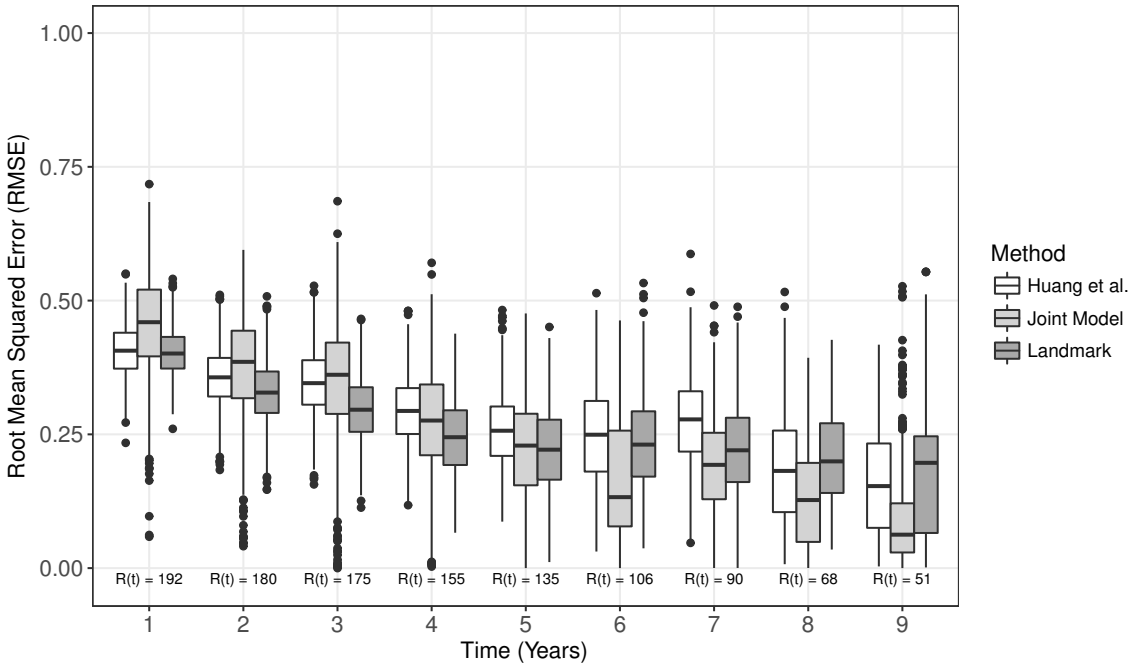


Figure 2.4. Root-mean-squared prediction error of $\pi(t + u = 10|t)$ for the landmark approach, joint modeling approach, and Huang et al. (2016) two-stage approach for times $t = 1, 2, \ldots, 9$ when $\sigma = 2$

ference relative to the landmark approach. While the results presented in Huang et al. (2016) are not separated by time t, the authors average (standard deviation) of the RMSE for the landmark approach, joint modeling approach, and their proposed method when υ in equation (2.13) is defined as $\upsilon = 0.8$ were 0.346 (0.021), 0.564 (0.020), and 0.344 (0.035), respectively. Given that the RMSE increases as υ increases, we would expect the RMSE computed in our simulation study to be greater than that of the previously reported values. However, examining Figure 2.3 and Table 2.2, it appears that our reported average RMSE for the joint modeling approach are lower for all time points $t > 1$. Similarly, our reported average RMSE for the landmark approach is lower than that reported by Huang et al. (2016) for all time points $t \in \{1, 2, \ldots, 9\}$. Lastly, we see that our reported average RMSE for the Huang et al. (2016) two-stage approach is less for times $t > 4$. The decrease in the average RMSE joint model may be attributed to the fact that in the original simulation study the

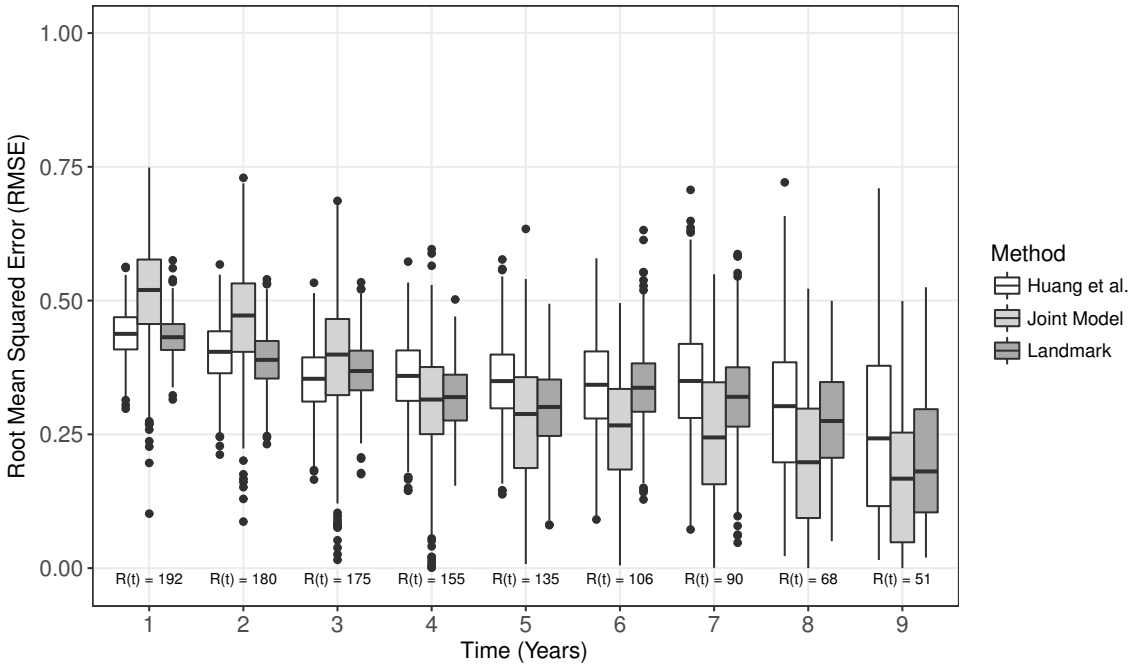


Figure 2.5. Root-mean-squared prediction error of $\pi(t + u = 10|t)$ for the landmark approach, joint modeling approach, and Huang et al. (2016) two-stage approach for times $t = 1, 2, \ldots, 9$ when $\sigma = 4$

authors used a nonparametric estimate of the baseline hazard function, which corresponds to the piecewise-constant model with knots positioned such that one event occurs between each knot, a practice known to lead to biased results.(Rizopoulos, 2012). However, we do notice that the standard deviations associated with the average RMSE for all three methods are larger than those reported by Huang et al. (2016). This results is to be expected given that we are considering prediction for a testing sample not involved in the training of the parameters associated with each of the three methods.

When comparing the performance of the three methods over time when the biologic variability is set at $\sigma = 2$, we see very little discrepancy between the joint modeling approach, landmark approach, and the Huang et al. (2016) two-stage approach, with no average RMSE being greater than two standard deviations apart between the three methods for any time point. When comparing the average RMSE

Table 2.2. Average (standard deviation) of root-mean-squared prediction error of the landmark approach, joint modeling approach, and two-stage approach proposed by Huang et al. (2016) for $\pi(T_i \geq 10 | T_i > t)$, $t = 1, 2, \ldots, 9$ for $\sigma = 0.6$.

Time (t)	R_t	Landmark	Joint Modeling	Huang et al. (2016)
$t = 1$	192	0.37 (0.05)	0.36 (0.09)	0.42 (0.06)
$t = 2$	180	0.29 (0.07)	0.19 (0.10)	0.40 (0.08)
$t = 3$	175	0.25 (0.08)	0.15 (0.09)	0.36 (0.10)
$t = 4$	155	0.20 (0.08)	0.10 (0.06)	0.36 (0.12)
$t = 5$	135	0.16 (0.09)	0.08 (0.06)	0.33 (0.14)
$t = 6$	106	0.15 (0.09)	0.06 (0.05)	0.34 (0.14)
$t = 7$	90	0.10 (0.07)	0.06 (0.05)	0.27 (0.17)
$t = 8$	68	0.09 (0.07)	0.05 (0.05)	0.20 (0.16)
$t = 9$	51	0.07 (0.06)	0.06 (0.07)	0.11 (0.13)

of the three methods between $\sigma = 0.6$ and $\sigma = 2$, we see a much larger increase in the average RMSE for the joint modeling approach than the landmark approach and two-stage approach proposed by Huang et al. (2016), increasing in the range of (0.07, 0.2). However, many of the average RMSE for the Huang et al. (2016) two-stage approach decrease with the exception of time $t = 9$, which increases by 5 points. Then, comparing the standard deviations of the RMSE for the three approaches when $\sigma = 0.6$ and $\sigma = 2$, we do not see a considerable increase.

Finally, when comparing the performance of the three methods over time when the biologic variability is specified to be $\sigma = 4$, we again see very little discrepancy between the three methods, with no average RMSE being greater than two standard deviations apart between the three methods for any time point. When comparing the average RMSE error for the three method between $\sigma = 2$ and $\sigma = 4$, we first see increases in the average RMSE for the landmark approach ranging from (0.03, 0.09). For the joint modeling approach we see changes ranging from (0.04, 0.09). Lastly, for the Huang et al. (2016) two-stage approach we see increases in average RMSE ranging from (0.00, 0.10). Additionally, when examining the standard deviations of the RMSE, we do not see a considerable increases between the three approach.

Table 2.3. Average (standard deviation) of root-mean-squared prediction error of the landmark approach, joint modeling approach, and two-stage approach proposed by Huang et al. (2016) for $\pi(T_i \geq 10|T_i > t)$, $t = 1, 2, \ldots, 9$ for $\sigma = 2$.

Time (t)	R_t	Landmark	Joint Modeling	Huang et al. (2016)
$t = 1$	192	0.40 (0.04)	0.45 (0.09)	0.41 (0.05)
$t = 2$	180	0.33 (0.06)	0.38 (0.10)	0.36 (0.06)
$t = 3$	175	0.30 (0.06)	0.35 (0.10)	0.35 (0.06)
$t = 4$	155	0.24 (0.07)	0.26 (0.12)	0.29 (0.06)
$t = 5$	135	0.22 (0.08)	0.22 (0.10)	0.26 (0.07)
$t = 6$	106	0.23 (0.09)	0.17 (0.11)	0.25 (0.09)
$t = 7$	90	0.22 (0.09)	0.19 (0.09)	0.28 (0.08)
$t = 8$	68	0.21 (0.09)	0.12 (0.08)	0.19 (0.10)
$t = 9$	51	0.17 (0.11)	0.10 (0.09)	0.16 (0.10)

Table 2.4. Average (standard deviation) of root-mean-squared prediction error of the landmark approach, joint modeling approach, and two-stage approach proposed by Huang et al. (2016) for $\pi(T_i \geq 10|T_i > t)$, $t = 1, 2, \ldots, 9$ for $\sigma = 4$.

Time (t)	R_t	Landmark	Joint Modeling	Huang et al. (2016)
$t = 1$	192	0.43 (0.04)	0.51 (0.09)	0.44 (0.04)
$t = 2$	180	0.39 (0.05)	0.47 (0.09)	0.40 (0.06)
$t = 3$	175	0.37 (0.05)	0.39 (0.10)	0.35 (0.06)
$t = 4$	155	0.32 (0.06)	0.31 (0.10)	0.36 (0.07)
$t = 5$	135	0.30 (0.08)	0.27 (0.12)	0.35 (0.08)
$t = 6$	106	0.34 (0.07)	0.26 (0.11)	0.34 (0.09)
$t = 7$	90	0.32 (0.09)	0.25 (0.12)	0.35 (0.10)
$t = 8$	68	0.28 (0.09)	0.20 (0.12)	0.29 (0.13)
$t = 9$	51	0.21 (0.11)	0.17 (0.13)	0.25 (0.15)

Based on the results for various degrees of biologic variability and the effect of increasing u in $\pi(T_i \geq t + u|T_i > t)$, we can see that the joint modeling approach performs ideally with minimal amounts of biologic variability (i.e, $\sigma = 0.6$). However, as the the biologic variability increases, the performance of the joint modeling approach becomes almost indistinguishable from that of the landmark approach and the Huang et al. (2016) two-stage approach. An interesting result of our simulation study is that the predictive performance of the Huang et al. (2016) approach does not differ considerably as biologic variability increases.

2.5 Advantages and Shortcomings of the Three Dynamic Prediction Approaches

We briefly mentioned the advantages and shortcomings of the landmark approach and the joint modeling approach as we introduced the respective models, but here we delve deeper into their respective flaws and benefits as well as the flaws and advantages associated with the Huang et al. (2016) two-stage approach.

First, the most glaring shortcoming of the landmark approach in the clinical setting is the unrealistic requirement that subjects have longitudinal biomarker measurements taken at the same landmark time point. As mentioned previously, a solution to this problem when subjects do not have measurements at exactly the same point is to employ a missing data technique. However, this has been shown to lead to additional bias in the estimated $\pi(T_i \geq t + u|T_i > t)$. Also, the landmark approach only considers biomarker measurements taken at the landmark time point and ignores all historical biomarker measurements, which may reveal additional evidence of the impending occurrence of an event-of-interest. Despite these flaws, the landmark approach provides an easily implementable option for dynamic prediction and, as we have shown, an equally effective method for dynamic prediction in the presence of large biologic variability.

Second, the joint modeling approach requires the specification of a longitudinal model for the biomarker measurements. As we previously mentioned, misspecification of this longitudinal model can result in biased estimates of $\pi(T_i \geq t + u|T_i > t)$. Additionally, while the large number of options provided to the user of the JM package in R offer the user the flexibility necessary to find the joint model which fits the data best, the amount of work necessary to obtain such a model requires considerable testing on the user's part as opposed to have a ready-to-go method for dynamic prediction straight out-of-the-box such as the landmark approach.

Lastly, while the Huang et al. (2016) two-stage approach does not specify a model for the biomarker trajectory, it does require specification of the time-varying coefficient model. The authors believed this consolation was a better option but the use of a large number of coefficients in the fractional polynomial model can lead to overfitting when the sample size may be small and require much more computational burden relative to both the joint modeling and landmark approach. Also, the prediction of the conditional survival probability at each time point t considers only the biomarker measurements at time t and at baseline, excluding biomarker history between those two time points. Finally, the Huang et al. (2016) two-stage approach does not have an R package or code to implement the method as do the two previously discussed methods. This requires any interested user or clinician to code the model from scratch, a time-consuming and less than ideal process given the other two methods.

2.6 Concluding Remarks and Future Works

As individualized medicine becomse common practice in research hospitals, dynamic prediction of longitudinal and time-to-event data has become a useful practice in adapting a patient's care to limit disease progression. We have conducted a simulation study to examine the predictive accuracy of the predicted conditional survival for the two commonly used dynamic prediction methods, the landmark approach and joint modeling approach, as well as the newly proposed Huang et al. (2016) two-stage approach. The simulation study conducted was constructed similarly to that in Huang et al. (2016), but instead examined the predictive accuracy of $\widehat{\pi}(T_i \geq t + u | T_i > t)$ for each method on a testing set patients after training on a training set of patients for differing values of u and differing degrees of biologic variability.

In our simulation study, we found that for biomarker measurement prone to minimal biologic variability, the joint modeling approach is most effective when the

longitudinal model of the biomarker is correctly specified. However, as biologic variability increases, the joint modeling approach, landmark approach, and the Huang et al. (2016) two-stage approach perform similarly.

Additionally, we have discussed the shortcomings and advantages of the three methods. If simplicity is key and data are all available at the landmark time point, the landmark approach is ideal. If the biologic variability of the longitudinal biomarker measurements is low, we encourage the use of the joint modeling approach with attention to correctly specifying the longitudinal model. If the biologic variability of the longitudinal measurements is high and data are available at the landmark time point, we suggest the use of the landmark approach given its simplicity and the absence of the need to specify the longitudinal biomarker measurements Huang et al. (2016). The Huang et al. (2016) two-stage approach does not appear to be advantageous in either situation given the computational complexity of the method relative to the landmark approach. However, we have only considered the scenario in which observations are all available at the landmark time point and the true longitudinal model is known. Future work should examine the performance of the Huang et al. (2016) two-stage approach when the longitudinal model is misspecified and data are not all measured at the landmark time point. The results previously reported by Huang et al. (2016) illustrated favorable performance of their proposed method, but given that their model was trained and tested on the same data during each iteration of the simulation study, their results may be representative of overfitting.

While the joint modeling approach is the method most investigated in dynamic prediction literature, development of a novel model which can incorporate history of the longitudinal biomarker measurements into the prediction of conditional survival while being robust to misspecification of the longitudinal model would be an invaluable contribution to the field of dynamic prediction. However, the inability to know longitudinal biomarker measurements for all time points from $[0, t]$, as well as future

values of the longitudinal biomarkers of interest, makes development of such a model difficult.

CHAPTER THREE

Inferring Cognitive Impairment from a Battery of Cognitive Tests at Baseline

3.1 Introduction

Multiple Sclerosis (MS) is an immune-mediated disease of the central nervous system affecting the brain, optic nerves, and spinal cord. Cognitive impairment has been well appreciated in the adult MS population, and recent research efforts have begun to document the presence of cognitive impairment in the pediatric MS population (Amato et al., 2008; Julian et al., 2013; MacAllister et al., 2005; Tan et al., 2017). Awareness of cognitive impairment in pediatric MS patients has led to an increase in research efforts aimed at identifying and treating cognitive impairment (MacAllister et al., 2005; Yeh et al., 2009).

The latent class status of cognitive impairment is inferred based on a patient's results from a battery, or collection, of multiple cognitive tests measuring various cognitive functions. Recently, the statistical methodology used to infer cognitive impairment status based on a given patient's results from a battery of cognitive tests has become a renewed topic of interest as cognitive impairment has become a outcome measure of interest in clinical trials (Huizenga et al., 2016). Much of the current efforts are aimed at controlling the Type I error in the presence of correlated cognitive tests in a given battery. Additionally, little has been done from the Bayesian perspective.

A characteristic of cognitive testing data that further complicates statistical analysis is the presence of censored data. Historically, censoring has been ignored because the most commonly used method converts test scores to 'pass'/'fail' variables, which is unaffected by censoring. However, as we investigate new methods for inferring cognitive impairment based on results from a battery of cognitive tests, accounting for censoring is beneficial to the estimation procedure.

Here we have investigated the use of a Bayesian multivariate normal mixture model to infer cognitive impairment from a battery of cognitive tests used for pediatric MS patients at the Pediatric Demyelinating Diseases Clinic of Children's Health Dallas with and without censoring. We first constructed a simulation study assuming independence among cognitive tests in a given battery and found that the Bayes procedure achieves a similar sensitivity and specificity relative to the commonly used methods. More importantly, we constructed a simulation based on the results from the analysis of the pediatric MS patients with correlated cognitive tests in the given battery and found that the Bayes procedure performed better than the commonly used methods in terms of both sensitivity and specificity. Finally, we examined the impact of censoring on the inferred cognitive status for the proposed model as well as the currently used methods. We found that despite accounting for censoring, parameter estimates in the Bayesian multivariate normal mixture model obtained when analyzing the censored dataset remain biased relative to the analysis of the uncensored dataset.

3.2 Background

A battery of cognitive tests consists of K cognitive tests administered to a patient during a single session. Let the random variable X_k denote the score on the k-th cognitive test in the battery, $k = 1, \ldots, K$, such that $\boldsymbol{X} = (X_1, X_2, \ldots, X_K)^\top$ is the K-dimensional random vector of cognitive tests scores. Additionally, the marginal distribution of X_k for a healthy (i.e., not cognitively impaired) patient is assumed to be normally distributed with known mean μ_k and standard deviation σ_k. Let Y_k be the centered and scaled score on the k-th cognitive test such that the marginal distribution of Y_k is assumed to be the standard normal distribution and $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_K)^\top$ is the K-dimensional random vector of standardized cognitive test scores. Going forth we will only discuss the vector $\boldsymbol{Y}$ as opposed to $\boldsymbol{X}$.

Consider a K-dimensional vector of data which are censored above at $\boldsymbol{b} = (b_1, b_2, \ldots, b_K)^\top$ and below at $\boldsymbol{a} = (a_1, a_2, \ldots, a_K)^\top$. Let the observed cognitive test results for the i-th subject on the k-th test be defined as

$$y_{ik}^* = y_{ik} \cdot I_{[a_k, b_k]}(y_{ik}) + a_k \cdot I_{(-\infty, a_k)}(y_{ik}) + b_k \cdot I_{(b_k, \infty)}(y_{ik}), \qquad \forall i, \forall k, \tag{3.1}$$

where $I_A(y_{ik})$ denotes the indicator function which equals 1 if $y_{ik} \in A$ and is 0 otherwise (Lee and Scott, 2012). Historically in cognitive test scoring, a patient's score is censored below and above at 3 standard deviations below and above the mean, respectively. That is, if y_{ik} has a marginal standard normal distribution, $a_k = -3$ and $b_k = 3$, for $k = 1, \ldots, K$, in equation (3.1). We can then define the random vector of potentially censored cognitive test scores as $\boldsymbol{Y}^* = (Y_1^*, Y_2^*, \ldots, Y_K^*)^\top$, such that

$$Y_k^* = \begin{cases} 3, & \text{if } Y_k > 3; \\ Y_k, & \text{if } Y_k \in [-3, 3]; \\ -3, & \text{if } Y_k < -3. \end{cases} \tag{3.2}$$

However, it is rare to encounter standardized test scores greater than three standard deviations above the mean but common to find test scores in cognitively impaired patients to fall below three standard deviations below the mean. Therefore, for the remainder of the text, when we refer to censoring we are referring to censoring below at three standard deviations below the mean (i.e., $Y_k < -3$).

3.2.1 Current Methods for Inferring Cognitive Impairment Based on a Battery of Cognitive Tests

3.2.1.1 Ingraham & Aiken (1996). Historically, the methods for inferring cognitive impairment based on the vector $\boldsymbol{Y}$ (or $\boldsymbol{Y}^*$) have relied upon first transforming $\boldsymbol{Y}$ (or $\boldsymbol{Y}^*$) into a binary 'pass'/'fail' variable based on score cut-off, c. Failure of a cognitive test is typically defined as scoring excessively low, so the value of c is generally negative and a patient is said to have 'failed' a given test if they scores lower

than c. Commonly used values of c in the neuropsychology literature are $c = -1$, $c = -1.5$, and $c = -2$ (Ingraham and Aiken, 1996). Let S be a random variable denoting the number of cognitive tests 'failed' such that s_i is the number of tests the i-th patient failed defined as

$$s_i = \sum_{k=1}^{K} I_{(-\infty,c)}(y_{ik}), \tag{3.3}$$

for some pre-defined c, where $I_A(y_{ik})$ denotes the indicator function which equals 1 if $y_{ik} \in A$ and is 0 otherwise. Ingraham and Aiken (1996) first proposed assuming S has a binomial distribution with K trials and probability of success equal to $\Phi(c)$, where $\Phi(\cdot)$ denotes the univariate standard normal cumulative distribution function (Ingraham and Aiken, 1996). Then, for a pre-defined significance level α, the ith patient is inferred to be cognitively impaired if $P(S \geq s_i) < \alpha$ based on the binomial cumulative distribution function, and not cognitively impaired otherwise. We will refer to this method as the IA Method for the remainder of this article for simplicity.

However, a well-recognized issue with IA Method, even by the authors themselves, is the assumption of independence among the tests in a cognitive battery. This assumption is often known to be invalid in various standard cognitive batteries, leading to an inflation of the probability of a Type I error (Berthelson et al., 2013).

An attractive feature of the IA Method is that it is not affected by the presence of censoring so long as the value of the cut-off, c, is greater than the value of a_k in equation (3.1).

3.2.1.2 P-value adjustments applied to the method proposed by Ingraham & Aiken (1996). As is common in the presence of multiple testing, p-value adjustment methods were incorporated into the IA Method. First, let the K-dimensional vector of p-values be defined as $\boldsymbol{p} = (p_1 = \Phi(Y_1), \ldots, p_K = \Phi(Y_K))^\top$, where $\Phi(\cdot)$ is the univariate standard normal cumulative distribution function. Then, through commonly used p-value adjustment methods including the Bonferroni correction, Holm

correction, one-step resampling, and step-down resampling, the vector of p-values are corrected for multiple testing and the resulting vector of adjusted p-values will be labled $\boldsymbol{p}^* = (p_1^*, \ldots, p_K^*)^\top$. For a pre-defined α, let S^* be the random variable denoting the number of cognitive tests for which the corrected p-value is below α such that for the i-th patient

$$s_i^* = \sum_{k=1}^{K} I_{[0,\alpha]}(p_{ik}^*), \tag{3.4}$$

where $I_A(y_{ik})$ denotes the indicator function which equals 1 if $y_{ik} \in A$ and is 0 otherwise. Then, similar to the IA Method, S^* is assumed to have a Binomial distribution with K trials and probability of success equal to α. Therefore, the i-th patient is inferred to be cognitively impaired if $P(S^* \geq s_i^*) < \alpha$ and not cognitively impaired otherwise.

The extensions of the IA Method using p-value adjustments may be affected by the presence of censoring given that the p-values computed prior to adjustment are based on Y_k^* as opposed to Y_k and $\Phi(Y_k) \leq \Phi(Y_k^*)$, for $k = 1, \ldots, K$.

3.2.1.3 Crawford et al. (2007). Crawford *et al.* (2007) proposed a novel method of inferring cognitive impairment from a battery of cognitive tests which did not rely on the Binomial distribution as does the method proposed by Ingraham and Aiken (1996) and its extensions (Crawford et al., 2007). The method proposed by Crawford *et al.* (2007) instead took advantage of the known correlation among tests in a standard battery of cognitive tests. Let Ω be a known $K \times K$ correlation matrix for a standard battery of K cognitive tests. Data were simulated from a multivariate normal distribution with a mean vector $\boldsymbol{\mu} = (0, \ldots, 0)^\top$ and correlation matrix Ω. Then, the probability a subject 'fails' at least k cognitive tests in a given battery was estimated to be the frequency of simulated data vectors with at least k simulated scores below a pre-defined value, $k = 1, \ldots, K$. If the number of tests a given subject 'fails' has a frequency of less than α in the simulated sample, the patient is classified

as cognitively impaired and not cognitively impaired otherwise. We will refer to the method proposed by Crawford *et al.* (2007) as the C Method for the remainder of the article for simplicity.

An issue that arises when using the C Method is that their approach is interested only in the number of tests below a pre-defined value, and not which tests are below a pre-defined value. This characteristic makes it possible to overestimate the probability of k tests being below a pre-defined value, especially if the k tests on which a patient scores below the pre-defined value are uncorrelated. Additionally, in order to implement this method, the correlation among the cognitive tests in the battery, $\boldsymbol{\Omega}$, must be known.

However, given that the C Method is interested only in the number of values falling below a given cut-off, it is not affected by the presence of censoring as long as the cut-off is greater than a_k in equation (3.1).

3.2.1.4 Huizenga et al. (2007). Huizenga *et al.* (2007) were one of the first to propose a multivariate approach which does not require the transformation of the vector $\boldsymbol{Y}$ (or $\boldsymbol{Y}^*$) into a vector of binary 'pass'/'fail' variables (Huizenga et al., 2007). However, the method proposed by Huizenga *et al.* (2007) requires a normative sample from healthy (i.e., not cognitively impaired) individuals. Let $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ be the results of a battery of cognitive tests from n healthy individuals and let $\boldsymbol{y}_{\text{new}}$ be the results from a battery of cognitive tests for a potentially cognitively impaired patient. Huizenga *et al.* (2007) then computed the T^2 test statistic as follows:

$$T^2 = (\boldsymbol{y}_{\text{new}} - \bar{\boldsymbol{y}})^\top \widehat{\Sigma}^{-1} (\boldsymbol{y}_{\text{new}} - \bar{\boldsymbol{y}}), \tag{3.5}$$

such that $\bar{\boldsymbol{y}} = \sum_{i=1}^{n} \boldsymbol{y}_i / n$ and $\widehat{\Sigma}$ is the $K \times K$ sample covariance matrix corresponding to the healthy sample. Then, given that $T^2 \sim \frac{(n^2-1)K}{n(n-K)} F_{K,n-K}$, the critical value corresponding to a desired α can be computed and patients whose scores from a battery of cognitive tests result in a T^2 value greater than the critical value are

inferred to be cognitively impaired and not cognitively impaired otherwise. We will refer to the method proposed by Huizenga *et al.* (2007) as the H Method for the remainder of the article for simplicity.

The T^2 statistic above is the T^2 statistic for testing the equality of vector means from two multivariate normal populations with equal covariance matrices, where $n_1 = n$ and $n_2 = 1$. Thus, the T^2 statistic above and its respective distribution operate under the assumption that $\boldsymbol{y}_{\text{new}}$ and the independent and identically distributed observations $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ are all normally distributed with a common covariance matrix, Σ , but potentially different means. A concern that must be addressed is if the covariance matrix of the cognitively impaired population of patients can be assumed to be equal to that of the healthy population. Additionally, this method requires that a reference sample of healthy individuals be available for comparison.

The H Method can potentially be affected by the presence of censoring. If $\boldsymbol{y}^*_{\text{new}}$ denotes the censored vector of a potentially cognitively impaired patient, it follows that $(\boldsymbol{y}_{\text{new}} - \bar{\boldsymbol{y}})^\top \widehat{\Sigma}^{-1} (\boldsymbol{y}_{\text{new}} - \bar{\boldsymbol{y}}) \geq (\boldsymbol{y}^*_{\text{new}} - \bar{\boldsymbol{y}})^\top \widehat{\Sigma}^{-1} (\boldsymbol{y}^*_{\text{new}} - \bar{\boldsymbol{y}})$. Hence, ignoring censoring can lead to underestimated values of T^2 and potential misclassificaton.

3.3 Multivariate Normal Mixture Model

Suppose that each of the $i = 1, \ldots, n$ pediatric MS patients in a given sample is sampled from either a cognitively impaired subpopulation or healthy (i.e., not cognitively impaired) subpopulation, with each latent subpopulation distribution belonging to the same parametric family with potentially different parameter values. Let z_i denote the latent class status for the i-th patient such that

$$z_i = \begin{cases} 1, & \text{if the } i\text{-th patient is from the healthy subpopulation;} \\ 2, & \text{if the } i\text{-th patient is from the cognitively impaired subpopulation.} \end{cases} \tag{3.6}$$

It follows that the distribution of the results from a battery of cognitive tests for the i-th patient, conditional on latent class status, is defined to be

$$\boldsymbol{y}_i|z_i \sim f(\boldsymbol{\theta}_{z_i}), \tag{3.7}$$

where $f(\boldsymbol{\theta}_{z_i})$ is a parametric distribution with parameters $\boldsymbol{\theta}_{z_i}$ that potentially differ for each subpopulation (Gelman et al., 2014). Then, let λ denote the proportion of the population of pediatric MS patients who are not cognitively impaired such that $\Pr(z_i = 1) = \lambda$ and let $(1 - \lambda)$ denote the proportion of the population of pediatric MS patients who are cognitively impaired such that $\Pr(z_i = 2) = 1 - \lambda$. The resulting likelihood obtained after marginalizing out the latent class status is defined to be

$$L(\boldsymbol{y}|\lambda, \boldsymbol{\theta}) = \lambda f(\boldsymbol{y}|\boldsymbol{\theta}_1) + (1 - \lambda) f(\boldsymbol{y}|\boldsymbol{\theta}_2), \tag{3.8}$$

which corresponds to a finite multivariate mixture model with two components, with component 1 assigned probability weight λ and component 2 assigned probability weight $(1 - \lambda)$ (Gelman et al., 2014).

For many cognitive tests, including the cognitive tests in which we are concerned, the marginal distribution of each cognitive test for a healthy (i.e., not cognitively impaired) patient is assumed to be normally distributed with known mean and standard deviation. That is, for the centered and scaled score on the kth cognitive test in a battery, $k = 1, \ldots, K$,

$$(y_{ik}|z_i = 1) \sim N(0, 1). \tag{3.9}$$

However, the marginal distribution of each cognitive test in a given battery for the cognitively impaired subpopulation is not known. Based on the assumed marginal distribution of each test in the healthy subpopulation shown in equation (3.9), we will begin by assuming a multivariate normal distribution for the joint distribution of the K cognitive test scores for a given battery for both subpopulations. The validity of this assumption will be examined upon analysis.

Let $\boldsymbol{Y}|z_i \sim \mathcal{N}_K(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})$, where $\mathcal{N}_K(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})$ is a K-dimensional normal distribution with mean vector $\boldsymbol{\mu}_{z_i}$ and covariance matrix $\boldsymbol{\Sigma}_{z_i}$. Instead of discussing the covariance matrix, we will consider a common correlation matrix for the two latent subpopulations, $\boldsymbol{\Omega}$, and the K-dimensional vector of standard deviations $\boldsymbol{\sigma}_{z_i} = (\sigma_{1,z_i}, \ldots, \sigma_{K,z_i})^\top$, where $\boldsymbol{\Sigma}_{z_i} = (\text{diag}(\boldsymbol{\sigma}_{z_i}))\boldsymbol{\Omega}(\text{diag}(\boldsymbol{\sigma}_{z_i}))$. That is,

$$f(\boldsymbol{y}|\boldsymbol{\theta}_{z_i}) = \frac{1}{(2\pi)^{K/2}|(\text{diag}(\boldsymbol{\sigma}_{z_i}))\boldsymbol{\Omega}(\text{diag}(\boldsymbol{\sigma}_{z_i}))|^{1/2}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_{z_i})^\top \left[(\text{diag}(\boldsymbol{\sigma}_{z_i}))\boldsymbol{\Omega}(\text{diag}(\boldsymbol{\sigma}_{z_i}))\right]^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_{z_i})\right\}, \quad (3.10)$$

where $\boldsymbol{\theta}_{z_i} = (\boldsymbol{\mu}_{z_i}, \boldsymbol{\sigma}_{z_i}, \boldsymbol{\Omega})$. From equation (3.9), given that $\mu_{k,z_i=1} = 0$ and $\sigma^2_{k,z_i=1} = 1$, for $k = 1, \ldots, K$, it follows that $\boldsymbol{\mu}_1 = (0, \ldots, 0)^\top$ and $\boldsymbol{\sigma}_1 = (1, \ldots, 1)^\top$.

Under the assumption that the $\boldsymbol{Y}|z_i = 1$ and $\boldsymbol{Y}|z_i = 2$ are of the same parametric family, there are two parameters which can vary relative to the healthy subpopulation: $\boldsymbol{\mu}_2$ and $\boldsymbol{\sigma}_2$. Therefore, relative to the healthy subpopulation, the cognitively impaired subpopulation can be shifted (i.e., $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$) and/or have differing variability (i.e., $\boldsymbol{\sigma}_1 \neq \boldsymbol{\sigma}_2$).

3.4 Methods

Data utilized in this study were collected under an IRB approved protocol of pediatric MS patients treated at the Pediatric Demyelinating Diseases Clinic of Children's Health Dallas.

The current battery of cognitive tests used in the Pediatric Demyelinating Diseases Clinic for MS patients is a non-standard battery of cognitive tests consisting of $K = 15$ cognitive tests, including: California Verbal Learning Test (CVLT) Total, CVLT Trial 1, CVLT Trial 5, CVLT Recognition Hits, CVLT Long Delay, Grooved Pegboard (Dominant Hand), Grooved Pegboard (Non-Dominant Hand), Digit Span, Beery Visual Motor Integration (VMI), Beery Visual Perception (VP), Trail Making Test A (TMTA), Trail Making Test B (TMTB), Symbol Search, Symbol Digit

Modality Test (SDMT), and the Delis-Kaplan Executive Function System (D-KEFS) Letter Fluency Test. Given that the current battery of cognitive tests is not a standard battery, correlation amongst the 15 tests is not available in a reference manual. Per the reference manuals for each cognitive test in the current battery, all cognitive tests scores are adjusted for age. Given that we are interested in the results from a given battery of cognitive tests for pediatric patients, it is important to note that TMTA and TMTB are not available for patients under the age of 9 and SDMT and the Letter Fluency Test are not available for patients under the age of 8.

For scores missing due to the score not being available for their given age group, we will assume that the missing data mechanism corresponding to these missing test scores is ignorable because the scores are age-adjusted. Additionally, due to administrative error some patients are missing additional test scores. However, this missingness is not believed to be of relevance to the score that the patient would have received and, thus, we will assume that the missing data mechanism is also ignorable.

There are two versions of the current dataset, a censored and an uncensored version. In the censored dataset, of the 15 cognitive tests in the current battery, seven of the tests are subject to censoring below at $a_k = -3$: Grooved Pegboard (Dominant), Grooved Pegboard (Non-Dominant), Beery VMI, Beery VP, TMTA, TMTB, and SDMT. In the uncensored dataset, while the true value of many of the censored test scores are known, the TMTA, TMTB, Beery VMI, and Beery VP may still be censored based on the lower limits on the scores available in their respective reference manuals or the participant not completing the task in the allotted amount of time. If a patient takes an excessive length of time to complete the TMTA and TMTB, the lowest score on the TMTA and TMTB was defined to be $Y_k = -12.67$ ($X_k = -90$) at the discretion of the examiner. Additionally, the lowest score in the reference manual for Beery VMI and Beery VP is $Y_k = -3.67$ ($X_k = 45$). Therefore, these tests are censored below at $a_k = -12.67$ and $a_k = -3.67$, respectively, in the uncensored

dataset. Despite the uncensored dataset potentially containing censored test scores, though to a much lesser extent, we will refer to this dataset as the uncensored dataset for convenience. Lastly, no test scores in the current dataset were greater than three standard deviations above the mean, so censoring above three standard deviations above the mean was not a concern.

For the i-th patient, let $\boldsymbol{y}_{i,\text{obs}}$ denote a $k_{i,\text{obs}}$-dimensional vector of observed and uncensored test scores, $\boldsymbol{y}_{i,\text{cens}}$ denote a $k_{i,\text{cens}}$-dimensional vector of censored test scores, and $\boldsymbol{y}_{i,\text{mis}}$ denote a $k_{i,\text{mis}}$-dimensional vector of missing test scores, such that $k_{i,\text{obs}} + k_{i,\text{cens}} + k_{i,\text{mis}} = K$. The missing data $\boldsymbol{y}_{i,\text{mis}}$ and censored data $\boldsymbol{y}_{i,\text{cens}}$ will be treated as unknown parameters in the Bayesian model and their values will be estimated based on the joint posterior distribution $p(\boldsymbol{y}_{i,\text{mis}}, \boldsymbol{y}_{i,\text{cens}}, \boldsymbol{\theta}_{z_i=1}, \boldsymbol{\theta}_{z_i=2}|\boldsymbol{y}_{obs})$.(Stan Development Team, 2017)

All analysis were performed using Stan in R (version 3.4.3) via the Rstan package (version 2.17.3) (R Core Team, 2016; Stan Development Team, 2016, 2017). Finite multivariate mixture models were fit with $J = 2$ components corresponding to the cognitively impaired and healthy subpopulations assuming a multivariate normal distribution for each subpopulation. All models were run with 5,000 warmup iterations and 20,000 post-warmup iterations. The model was initially run with three chains at different starting values to ensure convergence of each model to a stationary distribution. After determining convergence, each model was rerun with a single chain for the estimation of the joint posterior distribution of the unknown parameters, missing values, and censored values conditional on the observed data.

3.4.1 Priors

Due to the current battery of interest being a non-standard battery of cognitive tests, limited information is available regarding the correlation amongst tests. Additionally, limited information is available for the location and scale parameters

corresponding to the cognitively impaired subpopulation of pediatric MS patients. Therefore, weakly informative prior distributions will be used for most priors in each model to reflect our lack of knowledge.

While the proportion of pediatric MS patients suffering from cognitive impairment has been estimated to be approximately one-third, this estimate was obtained using the currently implemented methods to which we are comparing our method (Amato et al., 2008; Julian et al., 2013; MacAllister et al., 2005). Therefore, we will not use a strongly informative prior distribution for the distribution of λ which favors this estimate. We will specify a Beta(4, 4) prior distribution for λ to provide low probabilities that λ is close to zero or one.

The prior distributions specified for the healthy subpopulation (i.e., $\boldsymbol{Y}_i|z_i = 1$) and the cognitively impaired subpopulation (i.e., $\boldsymbol{Y}_i|z_i = 2$) are listed below. Due to the limited number of patients in the current sample, a common correlation matrix, $\boldsymbol{\Omega}$, was assumed for both subpopulations.

$$
\begin{aligned}
\boldsymbol{\Omega} &\sim \text{LKJ-Correlation Distribution}(\eta = 1) \\
\boldsymbol{\sigma}_1 &= (1, \ldots, 1)^\top \\
\boldsymbol{\sigma}_2 &\sim \text{Log-normal}(0, 1) \\
\boldsymbol{\mu}_1 &= (0, \ldots, 0)^\top \\
\boldsymbol{\mu}_2 &\sim \mathcal{N}\left((0, \ldots, 0)^\top, (\text{diag}(\boldsymbol{\sigma}_2))\boldsymbol{\Omega}(\text{diag}(\boldsymbol{\sigma}_2))\right)
\end{aligned}
$$

Per the marginal distribution of each tests shown in equation (3.9), the mean vector was specified to be $\boldsymbol{\mu}_1 = (0, \ldots, 0)^\top$ and the vector of standard deviations was specified to be $\boldsymbol{\sigma}_1 = (1, \ldots, 1)^\top$. The LKJ-correlation distribution with $\eta = 1$ prior distribution was chosen for the correlation matrix because when $\eta = 1$ the density is uniform over correlation matrices of order $K = 15$, while maintaining symmetry and positive definiteness (Stan Development Team, 2017; Lewandowski et al., 2009). Therefore, the LKJ-correlation distribution provides a weakly informative prior for

the correlations amongst the test to reflect our prior knowledge. A log-normal(0, 1) distribution was chosen as the prior distribution for the standard deviations of each of the K tests because the 2.5th and 97.5th percentiles of the log-normal(0, 1) distribution are 0.14 and 7.10, respectively, and given that the greatest range for any test in the censored dataset is 5.67 and the greatest range for any test in the uncensored dataset is 16.33, the log-normal(0, 1) provides a weakly informative prior. Finally, the prior distribution specified for $\boldsymbol{\mu}_2$ was chosen because had we chosen to specify conjugate prior distributions for the multivariate normal distribution, the prior distribution for $\boldsymbol{\mu}_2$ is $\mathcal{N}_K(\boldsymbol{\mu}_0, (\text{diag}(\boldsymbol{\sigma}_2))\boldsymbol{\Omega}(\text{diag}(\boldsymbol{\sigma}_2))/\kappa_0)$, where κ_0 is the prior number of measurements on the $(\text{diag}(\boldsymbol{\sigma}_2))\boldsymbol{\Omega}(\text{diag}(\boldsymbol{\sigma}_2))$ scale (Gelman et al., 2014). Therefore, given our lack of information regarding the mean of the K cognitive test scores obtained from the 'cognitively impaired', we specified $\boldsymbol{\mu}_0$ to be the zero vector and that we have 1 prior measurement on the $(\text{diag}(\boldsymbol{\sigma}_2))\boldsymbol{\Omega}(\text{diag}(\boldsymbol{\sigma}_2))$ scale, which is weakly informative.

3.4.2 Subpopulation Membership Estimate

The probability that a patient belongs to each subpopulation was estimated using the posterior estimates of the parameters $\boldsymbol{\theta}_{z_i}$ from the two subpopulations. Let $\pi_j(\boldsymbol{y}_i)$ denote the probability that the i-th patient belongs to the j-th subpopulation, $j = 1, 2$. It follows that

$$\begin{aligned}\pi_j(\boldsymbol{y}_i) &= P(z_i = j|\boldsymbol{y}_i, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &= \frac{\lambda_j f(\boldsymbol{y}_i|\boldsymbol{\theta}_j)}{\lambda f(\boldsymbol{y}|\boldsymbol{\theta}_1) + (1-\lambda) f(\boldsymbol{y}|\boldsymbol{\theta}_2)},\end{aligned}$$

for $j = 1, 2$, where $\lambda_2 = (1 - \lambda)$. A patient is then classified as cognitively impaired if $E[p(\pi_1|\boldsymbol{y}_i, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)] < E[p(\pi_2|\boldsymbol{y}_i, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)]$ and not cognitively impaired otherwise. Let n_1 denote the number of patient estimated to belong to the healthy subpopulation and let n_2 denote the number of patients estimate to belong to the cognitively impaired subpopulation.

3.4.3 Posterior Predictive Check

After estimating latent subpopulation membership, the squared Mahalanobis distance is to be computed based on the estimated parameters for respective subpopulation. Let d^2_{i,z_i} denote the squared Mahalanobis distance for the ith subject estimated to belong to the z_i subpopulation calculated as

$$d^2_{i,z_i}(\boldsymbol{y}_i) = (\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_{z_i})^\top \widehat{\boldsymbol{\Sigma}}^{-1}_{z_i}(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_{z_i}), \tag{3.11}$$

such that $\widehat{\boldsymbol{\mu}}_{z_i}$ is the estimated mean vector for the z_i subpopulation and $\widehat{\boldsymbol{\Sigma}}_{z_i}$ is the estimated covariance matrix for the z_i population. Therefore, there are n_1 squared Mahalanobis distances corresponding to the patients estimated to belong to the healthy subpopulation and n_2 squared Mahalanobis distances corresponding to the patients estimated to belong to the cognitively impaired subpopulation. Then, using the posterior predictive distribution for each subpopulation, n_1 and n_2 new patients are simulated from their respective subpopulations. Let $\widetilde{\boldsymbol{y}}_{i,z_i}$ denote the ith simulated patient from the posterior predictive distribution of the z_ith subpopulation, for $i = 1, \ldots, n_{z_i}$. Then, the squared Mahalanobis distance is computed for the simulated patients as in equations (3.11). Let the squared Mahalanobis distance evaluated for the i-th simulated patient from the z_i-th subpopulation be denoted as $d^2_{i,z_i}(\tilde{\boldsymbol{y}}_i)$

Let $d^2_{(i),z_i}(\boldsymbol{y})$ and $d^2_{(i),z_i}(\widetilde{\boldsymbol{y}})$ denote the ith ordered squared Mahalanobis distance for real patients and simulated patients, respectively, such that $d^2_{(i),z_i} \leq d^2_{(i+1),z_i}$, for $i = 1, \ldots, (n_{z_i} - 1)$. If the model assumptions previously made are valid, we would expect $d^2_{(i),z_i}(\boldsymbol{y}) \approx d^2_{(i),z_i}(\widetilde{\boldsymbol{y}})$. That is, if we plot the n_{z_i} points $(d^2_{(i),z_i}(\boldsymbol{y}), d^2_{(i),z_i}(\widetilde{\boldsymbol{y}}))$, the values should lie on the bisection of the quadrant.

Let $\widehat{d}^2_{(i),z_i}(\boldsymbol{y})$ denote the i-th ordered posterior mean of the squared Mahalanobis distance from the z_i-th subpopulation computed as

$$\widehat{d}^2_{(i),z_i}(\boldsymbol{y}) = \left(\widehat{E}[p(d^2_{i,z_i}|\boldsymbol{y}, \boldsymbol{\theta}_{z_i})]\right)_{(i)}, \tag{3.12}$$

where (i) need not equal i. Then, let $\widehat{d}^{\,2}_{(i),z_i}(\widetilde{\boldsymbol{y}}_i)$ denote the posterior mean for the i-th ordered squared Mahalanobis distance obtained from the posterior predictive distribution computed as

$$\widehat{d}^{\,2}_{(i),z_{(i)}}(\widetilde{\boldsymbol{y}}_i) = \widehat{E}[p(d^2_{(i),z_{(i)}}|\widetilde{\boldsymbol{y}}, \boldsymbol{\theta}_{z_{(i)}})]. \tag{3.13}$$

We will therefore estimate the ordered pair $(d^2_{(i),z_i}(\boldsymbol{y}), d^2_{(i),z_i}(\widetilde{\boldsymbol{y}}))$ using $(\widehat{d}^{\,2}_{(i),z_i}(\boldsymbol{y}), \widehat{d}^{\,2}_{(i),z_i}(\widetilde{\boldsymbol{y}}))$.

For those patients who have missing and/or censored test scores, the missing and/or censored test scores will be replaced with the simulated draw from the posterior predictive distribution of the missing and/or censored test scores given the observed test scores at each iteration.

3.4.4 Comparison to Currently Used Methods

The IA Method will be implemented on the patients' results with cut-off values $c = -1$, -1.5, and -2. A patient will be diagnosed if $P(S \geq s_i) \leq 0.05$. At an α level of 5%, given that there are $K = 15$ tests in the battery of cognitive tests, the value of s_i needed for a patient to be classified as 'cognitively impaired' using IA Method for $c = -1$, -1.5, and -2 are 6, 4, and 2, respectively. For those subjects with missing test scores, the number of trials in the binomial random variable are adjusted to the number of tests for which each subject has been scored. Censored values are used as they exist in the dataset.

The p-value adjustment methods that will be used are the Bonferroni, Holm, single-step multivariate permutation resampling, and step-down multivariate permutation resampling as discussed by Westfall and Young (1993) (Westfall and Young, 2002). Each individual test will be labelled as 'failed' if the adjusted p-value is less than 5%. Given that there are $K = 15$ tests, a patient will be classified as 'cognitively impaired' if they 'fail' at least 3 tests in the given battery. Again, for subjects with

missing observations, the number of trials in the binomial random variable will be adjusted accordingly. Censored values will be used as they are in the original dataset.

The C Method requires the correlation matrix corresponding to the tests in a given battery be known, as is the case for standard batteries of cognitive tests. However, the current battery is not a standard battery. Therefore, we will instead use the estimated correlation matrix for the healthy subpopulation obtained from the Bayesian multivariate mixture model to estimate the number of tests required that a patient 'fail' for them to be classified as cognitively impaired. The same cut-off score of -1.96 was used as in the original Crawford *et al.* (2007) article and it was determined that a patient needed to 'fail' at least 4 tests to be classified as 'cognitively impaired'. Again, for subjects with missing observations, the correlation matrix and mean vector will be marginalized to account for those tests observed and the number of tests a given patient is required to 'fail' to be considered cognitively impaired will be recomputed. Censored values will be used as they are in the original dataset.

The H Method requires a sample of known healthy patients. However, we do not have access to such a sample. Therefore, we will instead utilize the squared Mahalanobis distance of each patient using the parameters corresponding to the healthy subpopulation and a mean vector $\boldsymbol{\mu} = (0, \ldots, 0)^\top$. If the null distribution of $\boldsymbol{Y}$ is $\mathcal{N}_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the squared Mahalanobis distance is known to have an asymptotic chi-square distribution with $K = 15$ degrees-of-freedom. A subject will be classified as 'cognitively impaired' if their squared Mahalanobis distance is greater than $\chi^2_{15}(0.95) = 25.00$. The approach using the asymptotic distribution of the squared Mahalanobis distance operates under the same distributional assumptions as the H Method and the use of the F distribution corresponding to the T^2 shown in equation (3.5) converges to the chi-sqaure distribution with K degrees of freedom as the sample size of the healthy (i.e., not cognitively impaired) patients goes to infinity. The correlation matrix and mean vector will be marginalized to account for missing tests

for those patients with missing test scores. Censored values will remain censored as they are in the original dataset.

3.4.5 Leave-One-Out Cross Validation

In order to assess the potential for overfitting, leave-one-out cross validation was performed. Leave-one-out cross validation was used as opposed to larger testing samples due to the limited number of patients. For patient's in the training sample with missing and/or censored values, the missing and/or censored values are simulated from the posterior predictive distribution estimated from the $n-1$ training patients and then the posterior probability of subpopulation membership is estimated. The average in-sample and out-of-sample error will be computed relative to the subpopulation membership estimated using the full dataset.

3.5 Simulation

Simulations were constructed based on the results presented later in Section 3.6 to assess the ability of the Bayesian multivariate normal mixture model and the currently used methods to correctly infer a given patients subpopulation membership. Data were first simulated with $n = 45$ total patients, with $n_1 = 23$ patients sampled from the healthy subpopulation and $n_2 = 22$ patients simulated from the cognitively impaired subpopulation based on the estimated proportion of pediatric MS patients belonging to each subpopulation presented in Section 3.6. Data were simulated from four multivariate mixture distributions. Model 1 is meant to reflect the results from the actual dataset. The purpose of including Model 2 in our simulations is to examine the Bayesian multivariate normal mixture model's capabilities in the situation ideal for IA Method (i.e., uncorrelated tests in given battery). The purpose of Models 3 and 4 was to investigate the methods' performances with greater variability than specified by the multivariate normal model. The four multivariate mixture distributions from

which the data were simulated are shown below:

$$\begin{aligned}
\text{Model 1}: &\boldsymbol{Y}|z_i = 1 \sim \mathcal{N}_{15}(\mathbf{0}, \boldsymbol{\Omega}),\\
&\boldsymbol{Y}|z_i = 2 \sim \mathcal{N}_{15}(\boldsymbol{\mu}_2, (\text{diag}(\boldsymbol{\sigma}_2))\boldsymbol{\Omega}(\text{diag}(\boldsymbol{\sigma}_2))),\\
\text{Model 2}: &\boldsymbol{Y}|z_i = 1 \sim \mathcal{N}_{15}(\mathbf{0}, \boldsymbol{I}),\\
&\boldsymbol{Y}|z_i = 2 \sim \mathcal{N}_{15}(-1.2 \times \mathbf{1}, \boldsymbol{I});\\
\text{Model 3}: &\boldsymbol{Y}|z_i = 1 \sim \mathcal{T}_{15}(\nu = 4, \mathbf{0}, \boldsymbol{\Omega}),\\
&\boldsymbol{Y}|z_i = 2 \sim \mathcal{T}_{15}(\nu = 4, \boldsymbol{\mu}_2, (\text{diag}(\boldsymbol{\sigma}_2))\boldsymbol{\Omega}(\text{diag}(\boldsymbol{\sigma}_2)));\\
\text{Model 4}: &\boldsymbol{Y}|z_i = 1 \sim \mathcal{T}_{15}(\nu = 10, \mathbf{0}, \boldsymbol{\Omega}),\\
&\boldsymbol{Y}|z_i = 2 \sim \mathcal{T}_{15}(\nu = 10, \boldsymbol{\mu}_2, (\text{diag}(\boldsymbol{\sigma}_2))\boldsymbol{\Omega}(\text{diag}(\boldsymbol{\sigma}_2)));
\end{aligned}$$

such that $\boldsymbol{I}$ denotes the identity matrix, $\mathbf{1}$ denotes at 15-dimensional vectors of 1s, and $\mathbf{0}$ denotes a 15-dimensional vector of 0s. The values of $\boldsymbol{\Omega}$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\sigma}_2$ in Models 1, 3, and 4 are defined to be the posterior means of the respective parameters found in Table 3.4 under the censored and uncensored data columns and in Tables A.1, A.2, A.3, and A.4. The mean vector of $-1.2\times\mathbf{1}$ was chosen based on simulation to ensure that the IA Method for $c = -1$, -1.5, and -2 would have a power of at least 80% and α of at most 5% for n_1 and n_2 patients in the respective subpopulation. After simulating the n_1 and n_2 values in each model from their respective subpopulation distribution, censoring was applied using the method previous shown in equation (3.2). The missing values in the real dataset were reflected in the simulated datasets. For instance, if a subject was missing the TMTA and TMTB test scores in the real dataset and was inferred to belong to the healthy subpopulation, a subject in the simulated datasets from the healthy subpopulation would also be missing the TMTA and TMTB test scores. The diagnostic performance of all the Bayesian mixture models and the currently used methods are based on predictive accuracy of the known cognitive impairment status of the simulated patients. The simulation described above was then repeated to reflect the results presented in section 3.6 for the uncensored dataset. That is, 100

datasets were simulated with $n_1 = 24$ patients belonging to the healthy subpopulation and $n_2 = 21$ belonging to the cognitively impaired subpopulation, with the parameter values of $\boldsymbol{\Omega}$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\sigma}_2$ in Models 1, 2, and 3 are the defined to be the posterior means of the respective parameters found in 3.4 under the uncensored dataset columns and in Tables A.3 and A.4. All methods and priors described in section 3.4 were used in simulation with two exceptions. First, in the C Method, the true value of correlation matrix $\boldsymbol{\Omega}$ is used to conduct their procedures. Similarly, in the H Method, the true covariance of the healthy subpopulation was used. These changes were made to examine the ideal performance of both methods.

Table 3.1 provides the true positive rate (TPR) and false positive rate (FPR) for the multivariate normal mixture model and the currently used methods for each of the four simulation models in the presence of censoring such that a patient is positive if they are (or are inferred) to be cognitively impaired. From the results in Table 3.1, the multivariate normal mixture model has a large sensitivity for correctly inferring cognitive impairment across the four simulation models and exceptional specificity for the simulated data from Model 1 and Model 2, but the percentage of patients incorrectly inferred to be cognitively impaired increases as the degrees of freedom of the multivariate Student's t distribution decrease. The Holm, Bonferroni, and Single-Step p-value adjustments applied to the IA Method have perfect true negative rates but their ability to detect patients who are cognitively impaired in simulation is severely lacking. Conversely, the Step-Down p-value adjustment applied to the IA Method has a large sensitivity across all four simulation models, but the specificity of the method is very poor. The IA Method performs consistently across all four simulation models, with sensitivity increasing as c decreases and specificity decreasing as c decreases. More importantly, the multivariate normal mixture model has a high sensitivity with a much smaller range between the 97.5-th percentile and the 2.5-th percentile compared to the IA Method for all three values of c considered. However, for

the two simulated multivariate Student's t mixture models, the IA Method does have a greater specificity, with a smaller range between the 97.5-th percentile and the 2.5-th percentile of the FPR compared to the multivariate normal mixture model. The C Method performs exceptionally poorly on Model 2. However, the C Method maintains a constant sensitivity across the other three simulated models, which is less than that of the multivariate normal mixture model, but experiences a decrease in specificity for decreasing degrees of freedom in the multivariate Student's t distributions to a lesser extent than the multivariate normal mixture model. Finally, the H Method has a lower sensitivity than that of the multivariate normal mixture model on all four simulation models and experiences a much greater decrease in specificity for decreasing degrees of freedom in the simulated multivariate Student's t mixture models. When examining the results from the uncensored simulated data found in Table 3.2, we did not find any differences relative to those described above for the censored data with the exception of the sensitivity of the H Method. We see a much greater specificity for the H Method on Models 1, 3, and 4; achieving near equal sensitivity to the multivariate normal mixture model.

Finally, regarding the inferred subpopulation membership estimate, we can examine the agreement of the currently used methods and the Bayesian multivariate normal mixture model in the 100 simulated datasets from Model 1 for both the censored and uncensored scenario. The results are found in Table 3.3. From the results in Table 3.3 for the censored dataset, we see that for the Holm, Bonferroni, and Single-Step p-value adjustments to the IA Method result in a considerable disagreement in those inferred to be cognitively impaired by the multivariate normal mixture model. Conversely, the Step-Down p-value adjustment to the IA Method results in a large number of disagreements in those inferred to be not cognitively impaired by the multivariate normal mixture model. The IA Method for all three values of c results in similar number of disagreements in those inferred to be not cognitively im-

Table 3.1. True positive rate (TPR) and false positive rate (FPR) based on 100 simulated censored datasets of 45 patients; Positive = Cognitively Impaired, Negative = Not Cognitively Impaired

	Model 1		Model 2		Model 3		Model 4	
Method	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Multivariate Normal Mixture Model	0.95 (0.82, 1)	0.05 (0, 0.17)	0.98 (0.89, 1)	0.04 (0, 0.13)	0.87 (0.79, 0.92)	0.29 (0.05, 0.57)	0.87 (0.75, 0.92)	0.11 (0, 0.33)
Holm	0.43 (0.23, 0.64)	0 (0, 0)	0.08 (0, 0.18)	0 (0, 0)	0.46 (0.29, 0.62)	0.06 (0, 0.19)	0.42 (0.21, 0.62)	0.02 (0, 0.10)
Bonferroni	0.41 (0.2, 0.59)	0 (0, 0)	0.06 (0, 0.16)	0 (0, 0)	0.44 (0.27, 0.62)	0.05 (0, 0.17)	0.41 (0.21, 0.62)	0.01 (0, 0.10)
Single-Step	0.41 (0.23, 0.64)	0 (0, 0)	0.07 (0, 0.18)	0 (0, 0)	0.45 (0.27, 0.61)	0.05 (0, 0.17)	0.41 (0.21, 0.62)	0.02 (0, 0.10)
Step-Down	0.95 (0.86, 1)	0.31 (0.17, 0.48)	1 (1, 1)	0.2 (0.09, 0.3)	0.87 (0.79, 0.92)	0.44 (0.26, 0.62)	0.87 (0.79, 0.92)	0.38 (0.19, 0.57)
Ingraham & Aiken ($c = -1$)	0.81 (0.66, 0.95)	0.1 (0, 0.22)	0.95 (0.86, 1)	0.03 (0, 0.09)	0.74 (0.62, 0.88)	0.17 (0, 0.33)	0.74 (0.58, 0.88)	0.13 (0, 0.29)
Ingraham & Aiken ($c = -1.5$)	0.81 (0.64, 0.95)	0.06 (0, 0.17)	0.88 (0.75, 1)	0.02 (0, 0.09)	0.74 (0.58, 0.88)	0.17 (0, 0.33)	0.73 (0.58, 0.88)	0.11 (0, 0.24)
Ingraham & Aiken ($c = -2$)	0.89 (0.7, 1)	0.07 (0, 0.15)	0.86 (0.7, 0.95)	0.04 (0, 0.13)	0.83 (0.71, 0.92)	0.22 (0.05, 0.41)	0.82 (0.67, 0.92)	0.14 (0, 0.29)
Crawford	0.76 (0.59, 0.91)	0.02 (0, 0.09)	0.68 (0.5, 0.86)	0.01 (0, 0.07)	0.71 (0.56, 0.83)	0.14 (0, 0.31)	0.69 (0.50, 0.83)	0.08 (0, 0.22)
Huizenga	0.84 (0.66, 0.95)	0.05 (0, 0.13)	0.85 (0.73, 1)	0.05 (0.00, 0.13)	0.78 (0.58, 0.92)	0.36 (0.14, 0.57)	0.76 (0.62, 0.88)	0.21 (0.05, 0.43)

Table 3.2. True positive rate (TPR) and false positive rate (FPR) based on 100 simulated uncensored datasets of 45 patients; Positive = Cognitively Impaired, Negative = Not Cognitively Impaired

	Model 1		Model 2		Model 3		Model 4	
Method	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Multivariate Normal Mixture Model	0.98 (0.9, 1)	0.01 (0, 0.08)	0.98 (0.88, 1)	0.04 (0, 0.15)	0.97 (0.9, 1)	0.21 (0.02, 0.38)	0.98 (0.9, 1)	0.07 (0, 0.23)
Holm	0.59 (0.38, 0.79)	0 (0, 0)	0.08 (0, 0.17)	0 (0, 0)	0.63 (0.45, 0.81)	0.05 (0, 0.17)	0.60 (0.43, 0.81)	0.02 (0, 0.11)
Bonferroni	0.57 (0.38, 0.76)	0 (0, 0)	0.06 (0, 0.14)	0 (0, 0)	0.62 (0.43, 0.81)	0.05 (0, 0.17)	0.58 (0.43, 0.79)	0.02 (0, 0.11)
Single-Step	0.58 (0.38, 0.76)	0 (0, 0)	0.07 (0, 0.19)	0 (0, 0)	0.62 (0.43, 0.81)	0.05 (0, 0.17)	0.59 (0.43, 0.79)	0.02 (0, 0.08)
Step-Down	0.96 (0.86, 1)	0.31 (0.17, 0.46)	1 (1, 1)	0.2 (0.1, 0.31)	0.95 (0.86, 1)	0.4 (0.23, 0.58)	0.95 (0.86, 1)	0.35 (0.17, 0.5)
Ingraham & Aiken ($c = -1$)	0.82 (0.69, 0.95)	0.1 (0, 0.23)	0.95 (0.86, 1)	0.02 (0, 0.08)	0.82 (0.67, 0.95)	0.16 (0.02, 0.29)	0.81 (0.67, 0.95)	0.12 (0, 0.27)
Ingraham & Aiken ($c = -1.5$)	0.84 (0.71, 0.95)	0.06 (0, 0.17)	0.88 (0.76, 1)	0.02 (0, 0.08)	0.84 (0.71, 0.95)	0.16 (0.0, 0.29)	0.84 (0.67, 0.95)	0.1 (0, 0.23)
Ingraham & Aiken ($c = -2$)	0.92 (0.76, 1)	0.06 (0, 0.17)	0.86 (0.71, 0.95)	0.04 (0, 0.12)	0.93 (0.81, 1)	0.21 (0.02, 0.38)	0.92 (0.81, 1)	0.13 (0.02, 0.27)
Crawford	0.82 (0.67, 0.95)	0.03 (0, 0.12)	0.68 (0.5, 0.86)	0.01 (0, 0.06)	0.83 (0.67, 0.95)	0.13 (0, 0.25)	0.81 (0.67, 0.95)	0.07 (0, 0.21)
Huizenga	0.98 (0.9, 1)	0.05 (0.00, 0.15)	0.86 (0.76, 1)	0.05 (0.00, 0.12)	0.97 (0.88, 1)	0.33 (0.14, 0.54)	0.97 (0.88, 1)	0.20 (0.04, 0.38)

paired and those inferred to cognitively impaired, with a slightly larger number of disagreements in those inferred to be cognitively impaired, by the multivariate normal mixture model. The C Method results in a large number of disagreements among patients classified as impaired by the multivariate normal mixture model. Similarly, the H Method results in a large number of disagreements among patients classified as impaired by the multivariate normal mixture model. When examining the same results in the uncensored simulated dataset, we find many of the same trends discussed above but it is worth noting that the number of disagreements in patients inferred to cognitively impaired by the multivariate normal mixture model decreases for the Holm, Bonferroni, and Single-Step p-value adjustments to the IA Method, as well as the C Method and H Method.

3.6 Results

Results from a battery of cognitive tests were obtained for $n = 45$ pediatric MS patients. These patients underwent initial cognitive assessment via a battery of cognitive tests as part of an initial screening and independent of suspected cognitive impairment. Therefore, we will assume the sample of 45 patients represents a random sample of pediatric MS patients. The average age of patients was 15.08 years of age (range: 7 to 18 years of age) and 29 (64.4%) of the patients were female. A total of five patients were missing at least one cognitive test score in the current battery, with one patient missing a single cognitive test score, two patients missing two cognitive test scores, and two patients missing four cognitive test scores.

When examining the censored dataset, eleven patients had Grooved Pegboard (Dominant) scores censored, ten patients had Grooved Pegboard (Non-Dominant) scores censored, one patient had their Beery VMI score censored, one patient had their Beery VP score censored, eight had their TMTA score censored, 13 have their TMTB score censored, and one patient had their SDMT score censored. In the

Table 3.3. Posterior mean and 95% credible interval of the agreement between the Bayesian multivariate normal mixture model and the currently used methods on the inferred cognitive status based on the simulation of 100 censored and uncensored datasets containing 45 patients.

	Censored Simulated Dataset				Uncensored Simulated Dataset			
	Not Impaired		Impaired		Not Impaired		Impaired	
Method	Agree	Disagree	Agree	Disagree	Agree	Disagree	Agree	Disagree
Holm	23.03 (20, 25.52)	0.01 (0, 0)	9.36 (5, 14)	12.60 (8, 18.52)	24.07 (22, 26)	0.01 (0, 0)	12.34 (8, 16.52)	8.58 (4.47, 13)
Bonferroni	23.04 (20, 25.52)	0 (0, 0)	8.95 (4.47, 13)	13.01 (8, 18.52)	24.08 (22, 26)	0 (0, 0)	11.92 (8, 16)	9 (5, 13)
Single-Step	23.04 (20, 25.52)	0 (0, 0)	9.03 (5, 14)	12.93 (8, 18.05)	24.08 (22, 26)	0 (0, 0)	12.08 (8, 16)	8.84 (5, 13)
Step-Down	16.16 (12, 20)	6.88 (3, 11)	21.29 (18.48, 24.52)	0.67 (0, 3)	16.6 (13, 20.52)	7.48 (3.48, 11)	20.09 (18, 22)	0.83 (0, 3)
Ingraham & Aiken ($c = -1$)	20.99 (17, 24)	2.05 (0, 4.52)	17.97 (14, 21)	3.99 (1, 7.52)	21.71 (19, 24.52)	2.37 (0, 5)	17.36 (14, 20)	3.56 (1, 7)
Ingraham & Aiken ($c = -1.5$)	21.84 (18.48, 25)	1.20 (0, 4)	17.94 (14, 21.52)	4.02 (1, 7.52)	21.71 (19.48, 25)	1.45 (0, 4)	17.71 (14.48, 21)	3.21 (1, 6)
Ingraham & Aiken ($c = -2$)	21.72 (18.48, 25)	1.32 (0, 4)	19.79 (15, 23)	2.17 (0, 6)	22.52 (20, 25)	1.56 (0, 4)	19.32 (16, 21.52)	1.60 (0, 4)
Crawford	22.69 (20, 25.52)	0.35 (0, 2)	16.93 (13, 20.52)	5.03 (1, 9)	23.49 (21, 26)	0.59 (0, 2)	17.33 (14, 20.52)	3.59 (1, 7)
Huizenga	21.81 (19, 25)	1.23 (0, 4)	18.23 (15, 21)	3.73 (1, 8)	22.76 (20, 25)	1.32 (0, 4)	20.52 (19, 22)	0.4 (0, 2)

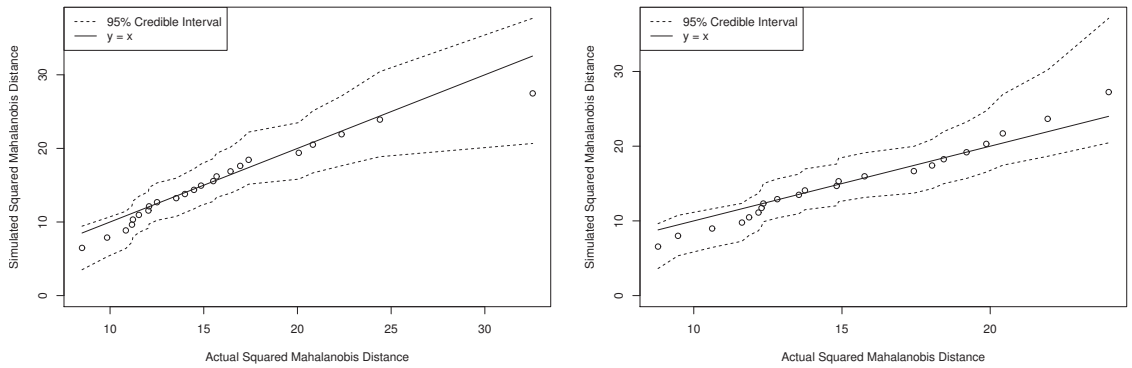


Figure 3.1. Posterior mean of the squared Mahalanobis distance based on the results from censored cognitive test scores for n_{z_i} patients inferred to be not cognitively versus the posterior mean and 95% credible interval of the squared Mahalanobis distance for n_{z_i} simulated patients; $n_1 = 23$, $n_2 = 22$.

uncensored dataset, one subject's score on the TMTB was below $Y_k = -12.67$, one subject's score on the Beery VMI was below $Y_k = -3.67$, and one subject's score on the Beery VP was below $Y_k = -3.67$. Therefore, the uncensored data still consisted of 3 censored cognitive test scores for which we are unable to obtain the true test score.

3.6.1 Censored Data

After ensuring model convergence via running multiple chains, the posterior predictive check was completed as discussed in section 3.4.3. The resulting plots of $(\widehat{d}^{\,2}_{(i),z_i}(\boldsymbol{y}), \widehat{d}^{\,2}_{(i),z_i}(\widetilde{\boldsymbol{y}}))$ for the two subpopulations are shown in Figure 3.1. There does not appear to be any indication of a poor fit for the multivariate normal model in Figure 3.1. Table 3.4 gives the posterior mean and 95% credible interval for the subpopulation membership proportion, as well as the mean and standard deviation vectors for the multivariate normal mixture model. The estimated correlation matrix can be found in the Appendix in Tables A.1 and A.2.

After performing the leave-one-out cross validation described in section 3.4.5, the average in-sample error ($\widehat{E}_{\text{in}}$) and out-of-sample error ($\widehat{E}_{\text{out}}$) were computed for the multivariate normal mixture model. The resulting average out-of-sample and in-

Table 3.4. Posterior mean and credible interval of parameters in the mixture of multivariate normal distributions.

	Censored Dataset		Uncensored Dataset	
Parameter[1]	Posterior Mean	95% Credible Interval	Posterior Mean	95% Credible Interval
λ	0.48	(0.33, 0.62)	0.52	(0.38, 0.66)
$\mu_{2,1}$	−1.04	(−1.42, −0.67)	−1.08	(−1.48, −0.69)
$\mu_{2,2}$	−0.64	(−0.94, −0.34)	−0.68	(−0.97, −0.39)
$\mu_{2,3}$	−1.02	(−1.44, −0.63)	−1.12	(−1.54, −0.70)
$\mu_{2,4}$	−0.75	(−1.16, −0.35)	−0.77	(−1.17, −0.36)
$\mu_{2,5}$	−0.53	(−1.02, −0.05)	−0.60	(−1.15, −0.06)
$\mu_{2,6}$	−2.51	(−3.85, −1.47)	−3.99	(−5.87, −2.13)
$\mu_{2,7}$	−2.62	(−3.74, −1.74)	−3.64	(−5.30, −2.01)
$\mu_{2,8}$	−0.73	(−1.06, −0.40)	−0.75	(−1.10, −0.38)
$\mu_{2,9}$	−1.46	(−1.89, −1.03)	−1.66	(−2.19, −1.13)
$\mu_{2,10}$	−0.92	(−1.39, −0.46)	−0.98	(−1.51, −0.47)
$\mu_{2,11}$	−1.85	(−2.86, −1.02)	−2.26	(−3.30, −1.24)
$\mu_{2,12}$	−3.20	(−4.91, −2.07)	−4.08	(−5.72, −2.47)
$\mu_{2,13}$	−1.35	(−1.81, −0.89)	−1.35	(−1.84, −0.86)
$\mu_{2,14}$	−0.83	(−1.35, −0.33)	−0.88	(−1.41, −0.37)
$\mu_{2,15}$	−0.76	(−1.35, −0.17)	−0.83	(−1.40, −0.21)
$\sigma_{2,1}$	0.91	(0.74, 1.14)	0.90	(0.72, 1.14)
$\sigma_{2,2}$	0.71	(0.55, 0.94)	0.66	(0.52, 0.85)
$\sigma_{2,3}$	0.94	(0.74, 1.22)	0.94	(0.74, 1.20)
$\sigma_{2,4}$	0.94	(0.73, 1.23)	0.91	(0.70, 1.19)
$\sigma_{2,5}$	1.13	(0.83, 1.54)	1.25	(0.91, 1.74)
$\sigma_{2,6}$	2.41	(1.61, 3.75)	4.39	(3.40, 5.78)
$\sigma_{2,7}$	1.98	(1.32, 3.02)	3.87	(2.99, 5.07)
$\sigma_{2,8}$	0.80	(0.62, 1.05)	0.80	(0.60, 1.07)
$\sigma_{2,9}$	1.01	(0.74, 1.38)	1.18	(0.85, 1.66)
$\sigma_{2,10}$	1.13	(0.83, 1.57)	1.16	(0.84, 1.66)
$\sigma_{2,11}$	1.96	(1.36, 2.91)	2.35	(1.77, 3.17)
$\sigma_{2,12}$	2.41	(1.50, 4.04)	3.71	(2.82, 4.97)
$\sigma_{2,13}$	1.11	(0.85, 1.47)	1.13	(0.84, 1.55)
$\sigma_{2,14}$	1.21	(0.92, 1.62)	1.22	(0.94, 1.63)
$\sigma_{2,15}$	1.37	(0.94, 1.89)	1.31	(0.84, 1.89)

[1] 1) CVLT Total, 2) CVLT Trial 1, 3) CVLT Trial 5, 4) CVLT Recognition Hits, 5) CVLT Long Delay, 6) Grooved Pegboard (Dominant Hand), 7) Grooved Pegboard (Non-Dominant Hand), 8) Digit Span, 9) Beery VMI, 10) Beery VP, 11) TMTA, 12) TMTB, 13) Symbol Search, 14) SDMT, 15) Letter Fluency Test.

sample error for the mixture of multivariate normal distributions was $\widehat{E}_{\text{out}} = 13.3\%$ and $\widehat{E}_{\text{in}} = 0.9\%$. Of the six patients whose inferred latent class status changed in the testing sample, four contained missing observations. Additionally, only two subjects' inferred cognitive status changed in the training sample, neither of which contained missing observations and only one of which contained a single censored cognitive test score.

Comparisons of inferred cognitive impaired status amongst the multivariate normal mixture model and the currently used methods are shown in Table 3.5. Based on the results presented in Table 3.5, relative to the inferred cognitive impairment status from multivariate normal mixture model, the p-value adjustments made to the method proposed by Ingraham and Aiken results in considerably different inferred cognitive status. Relative to the inferred cognitive status from the multivariate normal mixture model, the Holm, Bonferroni, and single-step method result in a greater number of patients being inferred as not cognitively impaired ($n_1 = 35$ (77.8%)). However, the step-down p-value adjustment infers a greater number of patients belonging to the cognitively impaired subpopulation ($n_2 = 29$ (64.4%)) relative to the multivariate normal mixture model ($n_2 = 22$ (48.9%)). Relative to the multivariate normal mixture model, these results suggest a greater agreement amongst the IA Method without p-value adjustments. However, there do exist differences amongst the inferred cognitive impairment status using the IA Method and the multivariate normal mixture model, with 9 (20%), 7 (15.6%), and 7(15.6%) inferred to belong to a different subpopulation for $c = -1$, $c = -1.5$, and $c = -2$, respectively. The C Method performs similarly to the IA Method relative to the multivariate normal mixture model, with a total of 8 (17.8%) patients whose inferred cognitive impairment status differs. Finally, the H Method infers a greater number of patients to belong to the healthy subpopulation (8) relative to the multivariate normal mixture model.

The results in Table 3.5 are similar to those obtained via simulation found in

Table 3.5. Agreement between the Bayesian multivariate normal mixture model and the currently used methods on the inferred cognitive status of the sample of pediatric MS patients using the censored or uncensored dataset.

	Censored Dataset				Uncensored Dataset			
	Not Impaired $n_1 = 23$		Impaired $n_2 = 22$		Not Impaired $n_1 = 24$		Impaired $n_2 = 21$	
Method	Agree	Disagree	Agree	Disagree	Agree	Disagree	Agree	Disagree
Holm	22 (95.7%)	1 (4.3%)	9 (40.9%)	13 (59.1%)	24 (100%)	0 (0%)	10 (47.6%)	11 (52.4%)
Bonferroni	22 (95.7%)	1 (4.3%)	9 (40.9%)	13 (59.1%)	24 (100%)	0 (0%)	9 (42.9%)	12 (57.1%)
Single-Step	22 (95.7%)	1 (4.3%)	9 (40.9%)	13 (59.1%)	24 (100%)	0 (0%)	9 (42.9%)	12 (57.1%)
Step-Down	13 (56.5%)	10 (43.5%)	19 (86.4%)	3 (13.6%)	14 (58.3%)	10 (41.7%)	19 (90.5%)	2 (9.5%)
Ingraham & Aiken ($c = -1$)	19 (82.6%)	4 (17.4%)	17 (77.3%)	5 (22.7%)	20 (83.3%)	4 (16.7%)	17 (81.0%)	4 (19.0%)
Ingraham & Aiken ($c = -1.5$)	21 (91.3%)	2 (8.7%)	17 (77.3%)	5 (22.7%)	23 (95.8%)	1 (4.2%)	18 (85.7%)	3 (14.3%)
Ingraham & Aiken ($c = -2$)	19 (82.6%)	4 (17.4%)	19 (86.4%)	3 (13.6%)	20 (83.3%)	4 (16.7%)	19 (90.5%)	2 (9.5%)
Crawford	21 (91.3%)	2 (8.7%)	16 (72.7%)	6 (27.3%)	23 (95.8%)	1 (4.2%)	17 (81.0%)	4 (19.0%)
Huizenga	23 (100%)	0 (0%)	14 (63.6%)	8 (36.4%)	24 (100%)	0 (0%)	20 (95.2%)	1 (4.8%)

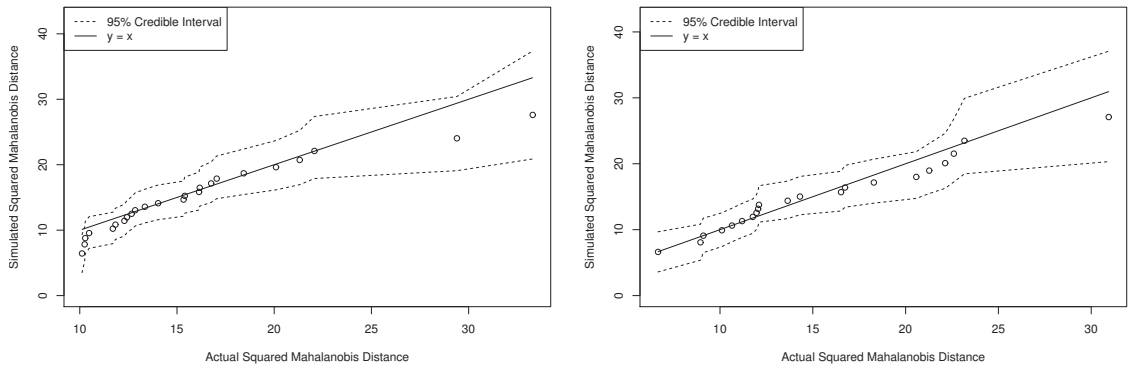


Figure 3.2. Posterior mean of the squared Mahalanobis distance based on the results from uncensored cognitive test scores for n_{z_i} patients inferred to be not cognitively versus the posterior mean and 95% credible interval of the squared Mahalanobis distance for n_{z_i} simulated patients; $n_1 = 24$, $n_2 = 21$.

Table 3.3 for the censored dataset. The only striking difference between the results in Table 3.5 and Table 3.3 for the censored dataset is that in the real dataset there is a single patient who is inferred to be impaired by the Holm, Bonferroni, and Single-Step p-value adjustment but inferred to be not cognitively impaired by the multivariate mixture model, while the average number of times this occurred in simulation is 0.1.

3.6.2 Uncensored Data

After ensuring model convergence via running multiple chains, the posterior predictive check was completed as discussed in section 3.4.3. The resulting plots of $(\widehat{d}^{\,2}_{(i),z_i}(\boldsymbol{y}), \widehat{d}^{\,2}_{(i),z_i}(\widetilde{\boldsymbol{y}}))$ for the mixture of multivariate normal distributions are shown in Figure 3.2. Based on the results shown in Figure 3.2, we do not see any obvious characteristics that demonstrate a poor fit for multivariate normal models we considered.

Table 3.4 gives the posterior mean and 95% credible interval for the subpopulation membership proportion, as well as the location and scale parameters for the multivariate normal mixture model. The estimated correlation matrix is found in TablesA.3 and A.4.

After performing the leave-one-out cross validation described in section 3.4.5,

the average in-sample error ($\widehat{E}_{\text{in}}$) and out-of-sample error ($\widehat{E}_{\text{out}}$) were computed for the multivariate normal mixture model. The resulting average out-of-sample and in-sample error for the mixture of multivariate normal distributions was $\widehat{E}_{\text{out}} = 15.6\%$ and $\widehat{E}_{\text{in}} = 2.3\%$. Of the seven patients whose inferred latent class status changed in the testing sample, four contained missing observations. Additionally, only four subjects' inferred cognitive status changed in the training sample, none of which contained missing or censored cognitive test scores.

Comparisons of inferred cognitive impaired status amongst the three mixture models and the currently used methods are shown in Table 3.5. The trends described previously in the comparison of the currently used methods to the mixture models for the censored dataset are very similar to those found in Table 3.5, with the exception of the H Method. Using the uncensored data, the H Method performs much more similar to the multivariate mixture models than previously observed in the censored dataset. This result is expected as the value of the T^2 statistic will be biased in the presence of censoring. The results in Table 3.5 are similar to those obtained via simulation found in Table 3.3 for the uncensored dataset. There is no obvious discrepancy between the results obtained via simulation and those obtaining using the real dataset.

3.6.3 Comparison of Censored and Uncensored Data Results

Three comparisons are of interest regarding the censored and uncensored datasets: 1) agreement in inferred subpopulation membership, 2) accuracy of posterior estimates of censored values, and 3) parameter estimates. When comparing changes in the inferred cognitive status between the censored dataset and the uncensored dataset for the multivariate normal mixture model, we see that two patients' diagnoses change from cognitively impaired to not cognitively impaired and one patient's diagnosis changes from not impaired to impaired. Additionally, we do see that the

Bonferroni and single-step p-value adjustments result in a single patient converting from impaired to not impaired when considering the uncensored dataset relative to the censored dataset. Then, as would be expected, no conversions occur when considering the IA Method and the C Method. Finally, as was observed previously, using the H Method we see an increase in the number of patients whose inferred cognitive impairment status changes to impaired when considering the uncensored data. Again, this result is not surprising given that the value of the T^2 statistic will be biased in the presence of censoring.

In the censored dataset there are a total of 42 censored test scores for which the true value of the censored test score is known. Given that the uncensored cognitive test scores are available, we can investigate the accuracy of the posterior estimates of these censored test scores. The 95% credible interval was obtained for the censored test scores and the predictive accuracy is measured based on the true value of the censored test score falling within the 95% credible interval. For the multivariate normal mixture model, 32 (76.2%) of the true values of the censored test scores fall in the 95% credible interval.

Based on the parameter estimates and their 95% credible intervals for the censored and uncensored datasets found in Table 3.4, respectively, we see that the large differences in the posterior estimates correspond to the location, shape, and scale parameters for the tests which have censored scores. Therefore, despite taking into account censoring when analyzing the censored dataset, the resulting parameter estimates remain biased. This result is to be expected given the increased amount of information provided by the uncensored test scores relative to the censored test scores. Based on the correlation estimates and their 95% credible intervals for the censored and uncensored datasets found in Tables A.1 and A.2 versus A.3 and A.4, respectively, there do not appear to be any striking discrepancies in the estimated correlations.

3.7 Discussion and Concluding Remarks

Evidence of cognitive impairment in the pediatric MS patient population has led to an increase in research efforts aimed at identifying and treating this symptom of the disease. However, in order for the latent class status of cognitive impairment to be used as a marker or outcome measure in clinical trials, a patient's inferred cognitive impairment status must be valid. Therefore, we propose the use of a Bayesian multivariate normal mixture model to infer cognitive impairment status to address the shortcomings of currently used methods.

The Bayesian multivariate normal mixture model and the currently used methods for inferring cognitive impairment were applied to a sample of pediatric MS patients treated at the Pediatric Demyelinating Diseases Clinic of Children's Health Dallas. In order to assess the validity of the Bayesian multivariate normal mixture model, data were simulated from a mixture of multivariate normal distributions corresponding to the estimated multivariate normal mixture model for the given sample. From this simulation, we found that the Bayesian multivariate normal mixture model performs exceptionally well in terms of sensitivity and specificity relative to the currently used methods. Additionally, the agreement in cognitive impairment status between the Bayesian multivariate normal mixture model and the currently used methods in the simulated dataset is similar to that of the agreement in the cognitive status between the Bayesian multivariate normal mixture model and the currently used methods in the actual dataset. This result provides further evidence of the effectiveness and superiority of the Bayesian multivariate normal mixture model relative to the currently used methods.

Furthermore, data were simulated from a multivariate mixture distribution assuming all tests in the given battery were independent in order to determine the effectiveness of the proposed method in a scenario ideal for the method proposed by Ingraham and Aiken (1996). As in the other simulation, the Bayesian multivariate

normal mixture model performed exceedingly well with sensitivity and specificity similar to the method proposed by Ingraham and Aiken (1996). This result illustrates the ability of the proposed model to effectively infer cognitive impairment even when the tests in a battery are independent.

Two other methods investigated require information regarding the correlation matrix of the tests in a given battery. However, without the correlation matrix being known or obtaining a sample of healthy subjects, it is not possible to implement either of these two methods. However, the Bayesian multivariate normal mixture model enables us to estimate the correlation matrix without the use of a sample of subjects known to be not cognitively impaired.

Finally, we were able to examine the effect of censoring on the inferred cognitive impairment status for the proposed method and the currently used method. Based on our results, censoring impacts the p-value adjustments to the method proposed by Ingraham and Aiken (1996), the method proposed by Huizenga *et al.* (2007), and the Bayesian multivariate normal mixture model. Regarding our proposed method, despite taking into account censoring, the parameter estimates corresponding to those tests which contain censored values remained biased. However, the estimated correlation matrix remained relatively unchanged in our analysis. Three patients' inferred cognitive impairment status changed between the analysis of the censored dataset and the analysis of the uncensored datasets. Given that the results of the uncensored simulated data are more similar to the results from the real uncensored data than the results from the censored simulated data are to the real censored data, as well as the persistence of bias in parameter estimates despite accounting for censoring, we suggest the use of uncensored data, if available, when inferring cognitive impairment status with the Bayesian multivariate mixture model.

We have provided a method for inferring cognitive impairment from a cohort of patients thought to contain healthy (i.e., not cognitively impaired) patients and

cognitively impaired patients with minimal prior information. Our results illustrate the effectiveness of the Bayesian multivariate normal mixture model in simulated situations meant to reflect what is observed in actual data. Additionally, we provided estimates for the location, scale, and correlation parameters for future studies and sample size calculations for clinical trials aimed at investigating cognitive impairment in the pediatric MS population.

Future efforts will focus on the application of the proposed method to larger sample sizes once available. Also, mixture models consisting of parametric families other than the multivariate normal distribution will be examined such as the multivariate Student's t distribution, multivariate skew normal distribution, and multivariate skew t distribution. These mixture models were examined using the current dataset but issues arose due to the limited number of patients in the current dataset. Lastly, an extension of the current work is to develop a longitudinal method to infer cognitive impairment status using the proposed model. However, a longitudinal method becomes complicated when considering the impact of the practice effect, a consequence of serial cognitive testing.

A limitation of the current study is due to the inability to know the true value of the latent class status. Therefore, we must rely upon simulation to verify the accuracy of our proposed method. Additionally, due to the rarity of pediatric MS, obtaining a large sample of patients is difficult. However, larger samples are needed for future studies which examine different mixture models as well as the potential for different correlation matrices for each subpopulation.

CHAPTER FOUR

Inferring Cognitive Impairment from a Battery of Cognitive Tests During Follow-Up

4.1 Introduction

When considering only baseline results from a battery of cognitive tests, the Bayesian multivariate normal mixture model worked well for inferring cognitive impairment from a battery of cognitive tests. However, a battery of cognitive tests is often administered multiple times during a patient's follow-up to assess potential changes in cognitive impairment status due to disease progression or therapeutic intervention. Therefore, we must have a method for inferring cognitive impairment from a battery of cognitive tests administered during follow-up.

An aspect of serial cognitive testing that complicates inferring cognitive status in repeated testing is the existence of the practice effect for various cognitive tests (Dikmen et al., 1999; Beglinger et al., 2005; Collie et al., 2003; Calamia et al., 2012). That is, a patient's cognitive test score may increase simply due to prior exposure to the test and the recency of such exposure. The practice effect can create an artificial perception that a patient's score may have increased when the score has stayed the same or, even worse, decreased. Therefore, we must account for the practice effect in order to correctly infer a patient's cognitive status in repeated cognitive testing.

We have proposed the use of a Bayesian continuous-time mixed hidden Markov model to infer cognitive impairment for serial administrations of a battery of cognitive tests. Limited longitudinal data exist currently from the battery of cognitive tests discussed previously in the pediatric MS population. Therefore, in preparation for such data, we have designed two simulated scenarios and have tested the proposed model's performance by examining the accuracy of the posterior estimates and classification of the simulated patients' cognitive impairment status.

4.2 Continuous-Time Hidden Markov Model

The application of the Bayesian multivariate mixture presented in Chapter Three to results from a battery of cognitive tests during follow-up would require the assumption of independence among cognitive test scores obtained over time within a given patient as well as the cognitive impairment status over time within a given patient. A relaxation of these independence assumptions can be achieved through the use of a hidden Markov model (HMM). A HMM is similar to the finite mixture model previously discussed in that the observed data are generated from a distribution dependent upon an underlying, unobserved Markov process, with the marginal distribution of a HMM being a finite mixture model (Zucchini et al., 2016).

The most basic, time-homogeneous HMM incorporates the correlation among the cognitive status during a patient's disease course. For the ith patient, $i = 1, 2, \ldots, n$, let $\boldsymbol{Y}_{i\ell}$ denote the observed results from a battery of cognitive tests measured at time point $t_{i\ell}$, $\ell = 1, 2, \ldots, m_i$, where m_i denotes the number of longitudinal measurements for the ith patient. Additionally, let $t_{i\ell} \in \mathbb{R}^+$ denote the time from baseline at which the battery of cognitive tests is administered, such that $t_{i1} = 0$. Extending the notation previously used to account for serial cognitive testing, let $Z_{i\ell}$ denote the random variable corresponding to the ith patient's cognitive status at time $t_{i\ell}$, $i = 1, 2, \ldots, n$, $\ell = 1, 2, \ldots, m_i$. In a HMM, we define

$$P(Z_{i\ell}|Z_{i(\ell-1)}, Z_{i(\ell-2)}, \ldots, Z_{i1}) = P(Z_{i\ell}|Z_{i(\ell-1)}),$$

and

$$P(\boldsymbol{Y}_{i\ell}|\boldsymbol{Y}_{i(\ell-1)}, \boldsymbol{Y}_{i(\ell-2)}, \ldots, \boldsymbol{Y}_{i1}, Z_{i\ell}, Z_{i(\ell-2)}, \ldots, Z_{i1}) = P(\boldsymbol{Y}_{i\ell}|Z_{i\ell}). \tag{4.1}$$

Based on equation (4.1), we see that the distribution of $\boldsymbol{Y}_{i\ell}$ depends only on the unobserved state of the Markov process at time $t_{i\ell}$, which indicates that we are still assuming $\boldsymbol{Y}_{i\ell}$ is independent of previous results from a battery of cognitive tests. We will address a method of incorporating dependence among the results from a battery

of cognitive tests from a single patient after introducing additional notation associated with the HMM.

Additionally, we need to define an important quantity of interest in the HMM, the transition probability matrix $\Gamma(t_{i\ell} - t_{i(\ell-1)})$. The transition probability matrix contains as its j, j'-th element the conditional probabilities of $Z_{i\ell} = j$ given $Z_{i(\ell-1)} = j'$; that is,

$$\Gamma(t_{i\ell} - t_{i(\ell-1)})_{j,j'} = P(Z_{i\ell} = j' | Z_{i(\ell-1)} = j). \tag{4.2}$$

Assuming that the Markov process is time-homogeneous, equation (4.2) simplifies to

$$\Gamma(t_{i\ell} - t_{i(\ell-1)})_{j,j'} = P(Z_i(t_{i\ell} - t_{i(\ell-1)}) = j' | Z_i(0) = j), \tag{4.3}$$

where $Z_i(t_{i\ell} - t_{i(\ell-1)})$ denotes the latent class status evaluated at time $(t_{i\ell} - t_{i(\ell-1)})$ from baseline, $Z_i(0)$. That is, the transition probability matrix shown in equation (4.3) does not depend upon $t_{i\ell}$ and $t_{i(\ell-1)}$, only the difference $(t_{i\ell} - t_{i(\ell-1)})$. In the scenario in which we are interested, $z_{i\ell}$ can assume two values such that $z_{i\ell} = 1$ if a patient is healthy (i.e., not cognitively impaired) and $z_{i\ell} = 2$ if a patient is cognitively impaired. Therefore, $j = 1, 2$ and $j' = 1, 2$.

For a continuous-time HMM under the assumption of a time-homogeneous Markov process, we can define the transition probability matrix as

$$\Gamma(t_{i\ell} - t_{i(\ell-1)}) = \text{Exp}[(t_{i\ell} - t_{i(\ell-1)})Q],$$

where Exp denotes the matrix exponential function, not the element-wise exponential function, and Q is a matrix of transition intensities which, for a two-state Markov process, can be defined as

$$Q = \begin{bmatrix} -q_1 & q_1 \\ q_2 & -q_2 \end{bmatrix}, \tag{4.4}$$

such that q_1, $q_2 \in \mathbb{R}^+$ (Jackson et al., 2011).

Define $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^\top$ to be the vector of component membership probabilities at t_{i1} and $P(\boldsymbol{y}_{i\ell}) = \text{diag}(f_1(\boldsymbol{y}_{i\ell}), f_2(\boldsymbol{y}_{i\ell}))$, such that $f_1(\boldsymbol{y}_{i\ell}) = P(\boldsymbol{Y}_{i\ell} = \boldsymbol{y}_{i\ell}|Z_{i\ell} = 1)$ and $f_2(\boldsymbol{y}_{i\ell}) = P(\boldsymbol{Y}_{i\ell} = \boldsymbol{y}_{i\ell}|Z_{i\ell} = 2)$. The likelihood for the ith patient can then be defined as

$$L_i = \boldsymbol{\lambda}^\top P(\boldsymbol{y}_{i1})\Gamma(t_{i2} - t_{i1})P(\boldsymbol{y}_{i2}) \cdots \Gamma(t_{im_i} - t_{i(m_i-1)})P(\boldsymbol{y}_{im_i})\mathbf{1}, \tag{4.5}$$

such that $\mathbf{1} = (1, 1)^\top$.

The introduction of $f_1(\boldsymbol{y}_{i\ell})$ and $f_2(\boldsymbol{y}_{i\ell})$ enable us to now present a method for accounting for correlation among results from a battery of cognitive tests from a given patient. We can specify $f_1(\boldsymbol{y}_{i\ell})$ and $f_2(\boldsymbol{y}_{i\ell})$ to be functions of fixed effects and random effects at the patient level. So-called mixed HMM have previously been investigated as a method for accounting for intra-subject correlation of the observed data given the latent class status (Van Montfort et al., 2010).

Two quantities of interest that must be defined are the forward probability j-dimensional vector $\boldsymbol{\alpha}_{i\ell}$ and the backwards probability j-dimensional vector $\boldsymbol{\beta}_{i\ell}$, $j = 1, 2, \ldots, J$. The vector $\boldsymbol{\alpha}_{i\ell}$ for the ith patient at the ℓth time point is defined as

$$\boldsymbol{\alpha}_{i\ell} = \boldsymbol{\lambda}P(\boldsymbol{y}_{i1})\Gamma(t_{i2} - t_{i1})P(\boldsymbol{y}_{i2}) \cdots \Gamma(t_{i\ell} - t_{i(\ell-1)})P(\boldsymbol{y}_{i\ell}), \tag{4.6}$$

such that the j-th element of $\boldsymbol{\alpha}_{i\ell}$ is $\boldsymbol{\alpha}_{i\ell}(j) = P(\boldsymbol{Y}_{i1} = \boldsymbol{y}_{i1}, \ldots, \boldsymbol{Y}_{i\ell} = \boldsymbol{y}_{i\ell}, Z_{i\ell} = j)$. It follows from equation (4.6) that

$$\boldsymbol{\alpha}_{i\ell} = \boldsymbol{\alpha}_{i(\ell-1)}\Gamma(t_{i\ell} - t_{i(\ell-1)})P(\boldsymbol{y}_{i\ell}), \tag{4.7}$$

and

$$L_i = \boldsymbol{\alpha}_{im_i}\mathbf{1}.$$

The vector $\boldsymbol{\beta}_{i\ell}$ is then defined to be

$$\boldsymbol{\beta}_{i\ell} = \Gamma(t_{i(\ell+1)} - t_{i\ell})P(\boldsymbol{y}_{i(\ell+1)}) \cdots \Gamma(t_{im_i} - t_{i(m_i-1)})P(\boldsymbol{y}_{im_i})\mathbf{1}, \tag{4.8}$$

where the j-th element of $\boldsymbol{\beta}_{i\ell}$ is $\boldsymbol{\beta}_{i\ell}(j) = P(\boldsymbol{Y}_{i(\ell+1)} = \boldsymbol{y}_{i(\ell+1)}, \ldots, \boldsymbol{Y}_{im_i} = \boldsymbol{y}_{im_i} | Z_{i\ell} = j)$ and $\boldsymbol{\beta}_{im_i} = \mathbf{1}$. It can be seen from equations (4.5), (4.6), and (4.8) that

$$\boldsymbol{\alpha}_{i\ell}(j)\boldsymbol{\beta}_{i\ell}(j) = P(\boldsymbol{Y}_{i1} = \boldsymbol{y}_{i1}, \ldots, \boldsymbol{Y}_{im_i} = \boldsymbol{y}_{im_i}, Z_{i\ell} = j),$$

where $\boldsymbol{\alpha}_{i\ell}(j)$ corresponds to the forward probability if $z_{i\ell} = j$, $j = 1, 2$, and

$$\boldsymbol{\alpha}_{i\ell}\boldsymbol{\beta}_{i\ell} = L_i.$$

Finally, we can compute the conditional probability

$$P(Z_{i\ell} = j | \boldsymbol{Y}_{i1} = \boldsymbol{y}_{i1}, \ldots, \boldsymbol{Y}_{im_i} = \boldsymbol{y}_{im_i}) = \boldsymbol{\alpha}_{i\ell}(j)\boldsymbol{\beta}_{i\ell}(j)/L_i. \tag{4.9}$$

The resulting quantity in equation (4.9) will be our quantity of interest when inferring cognitive impairment in serial cognitive testing.

4.3 Accounting for the Practice Effect in Serial Cognitive Testing

The practice effect has been well documented in serial cognitive testing for various cognitive tests, as well as the necessity of accounting for such an effect when analyzing serial cognitive testing results (Dikmen et al., 1999; Beglinger et al., 2005; Collie et al., 2003; Calamia et al., 2012). However, focus has generally been on incorporating the practice effect when examining changes over time in the scores $\boldsymbol{Y}_{ij}$. To our knowledge, the potential impact of the practice effect on the inferred cognitive status of a particular patient has not been investigated or incorporated into any methods. Therefore, we seek to add a covariate in the mixed HMM capable of accounting for the practice effect.

It has been well documented that the practice effect decays as the difference $t_{ij} - t_{i(j-1)}$ increases (Dikmen et al., 1999; Beglinger et al., 2005; Collie et al., 2003; Calamia et al., 2012). Therefore, we chose to incorporate a method proposed by Settles and Meeder (2016) termed half-life regression. Half-life regression is based on

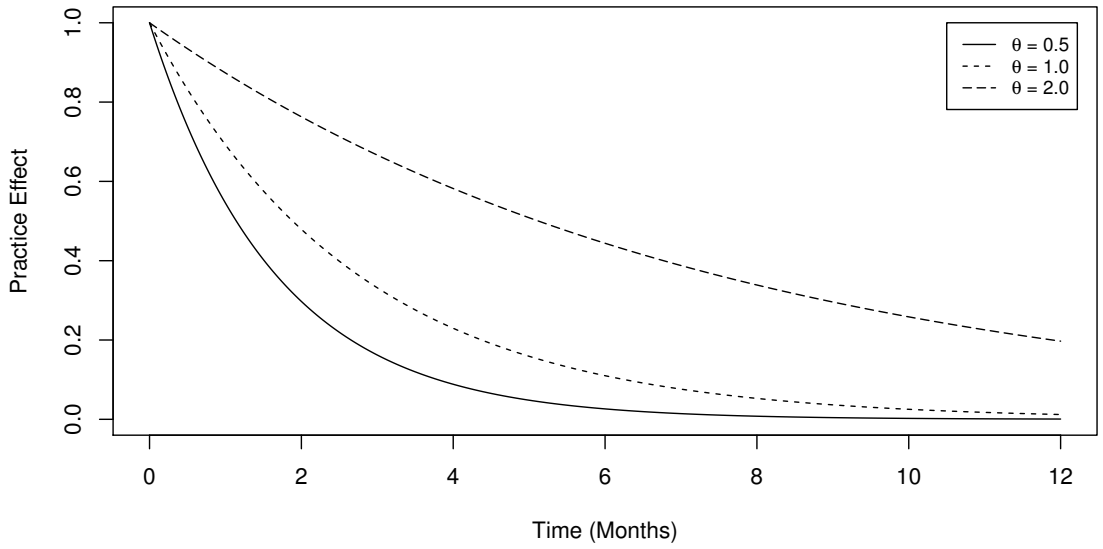


Figure 4.1. Practice effect as estimated by $\beta \exp\{-(t_{i\ell} - t_{i(\ell-1)})/\exp[\theta]\}$, such that $\beta = 1$ and $\theta = 0.5$, 1, and 2, for values of $(t_{i\ell} - t_{i(\ell-1)}) \in [0, 12]$.

the forgetting curve proposed by Ebbinghaus (2013), defined as

$$p = \exp\{-(t_{i\ell} - t_{i(\ell-1)})/h\}, \tag{4.10}$$

where p is a percentile of recall and h is the half-life. Settles and Meeder (2016) proposed defining the quantity $\widehat{h} = \exp\{\boldsymbol{\Theta}\boldsymbol{X}\}$, where $\boldsymbol{\Theta}$ are the parameters associated with the covariates $\boldsymbol{X}$. This definition of $\widehat{h}$ enables the incorporation of covariates into the model that impact the half-life in equation (4.10).

As was mentioned, the forgetting curve refers to percent recall. For example, the number of words learned and recalled after a delay. However, we are interested in applying this equation to the percentage of points increased from baseline due to the practice effect. Therefore, we will define a new quantity based on the half-life regression equation proposed by Settles and Meeder (2016), defined as

$$\beta p = \beta \exp\{-(t_{i\ell} - t_{i(\ell-1)})/\exp[\boldsymbol{\Theta}\boldsymbol{X}]\}, \tag{4.11}$$

where β is the number of points a patient's score would increase if tested immediate (i.e., $(t_{i\ell} - t_{i(\ell-1)}) = 0$) due to the practice effect. We will incorporate this parameter in

the distributions $f_1(\boldsymbol{y}_{i\ell})$ and $f_2(\boldsymbol{y}_{i\ell})$, with potentially subject-specific random effects included as β_i. A plot of a simple form of equation (4.11) in which $\boldsymbol{\Theta} = \theta \in \mathbb{R}$ and $\boldsymbol{X} = 1$ evaluated at three value of θ and over $(t_{i\ell} - t_{i(\ell-1)}) \in [0, 12]$ can be found in Figure 4.1.

4.4 Method

4.4.1 Data Generation

To examine the ability of the continuous-time mixed HMM, we have constructed two scenarios meant to reflect a subset of the results presented in Chapter Three. In the Scenario 1, we will simulate $n = 100$ patients with results from a single cognitive test at baseline, $t_{i1} = 0$, and four additional testings after baseline irregularly spaced in the interval $t_{i\ell} \in [2, 30]$ months such that $t_{i(\ell-1)} < t_{i\ell}$, $\ell = 2, \ldots, 5$. Therefore, there are a total of $m_i = 5$ measurements per patient. The transition intensities of the matrix Q in equation (4.4) are defined to be $q_1 = 0.6$ and $q_2 = 0.4$. As has been discussed, we will define two latent subpopulations where $z_{i\ell} = 1$ if a patient is healthy (i.e., not cognitively impaired) and $z_{i\ell} = 2$ if a patient is cognitive impaired. We will then define the vector $\boldsymbol{\lambda} = (0.5, 0.5)$, which is approximately the subpopulation membership estimates obtained in Chapter Three, and the latent class status for the ith patient at baseline will be simulated by randomly sampling W_{i1} from a $Bernoulli(p = 0.5)$ and defining $z_{i1} = w_{i1} + 1$. The latent class status for the ith patient at times $t_{i2}, \ldots, t_{im_i}$ are then determined by randomly sampling $W_{i\ell}$ from a $Bernoulli(p_{i\ell})$, such that $p_{i\ell}$ is defined as $P(Z_{i\ell} = 2 | Z_{i(\ell-1)} = z_{i(\ell-1)})$ computed using the $z_{i(\ell-1)}$, 2-nd element of $\Gamma(t_{i\ell} - t_{i(\ell-1)})$, and $z_{i\ell} = w_{i\ell} + 1$, $\ell = 2, \ldots, m_i$.

Lastly, for the two scenarios we will define $Y_{i\ell k} | z_{i\ell} = j$ to be the scores for the ith patient at time $t_{i\ell}$ on the kth test, $k = 1, \ldots, K$, in the cognitive battery conditional on the latent class status $z_{i\ell}$ at time $t_{i\ell}$. We will define the probability

density function corresponding to the latent subpopulation $z_{i\ell}$ to be the mixed effects model shown below:

$$(y_{i\ell k}|z_{i\ell}=j)=\begin{cases}\left(\beta_{0k}^{(j)}+b_{0ik}^{(j)}\right)+\epsilon_{i\ell k}^{(j)}, & \text{if } t_{i1};\\ \left(\beta_{0k}^{(j)}+b_{0ik}^{(j)}\right)+\left(\beta_{1k}^{(j)}+b_{1ik}^{(j)}\right)\exp\left\{-\frac{(t_{i\ell}-t_{i(\ell-1))}}{\exp\left[\theta_k^{(j)}\right]}\right\}+\epsilon_{i\ell k}^{(j)}, & \text{otherwise.}\end{cases}$$

such that $\beta_{0k}^{(j)}$ and $\beta_{1k}^{(j)}$ are the fixed effects corresponding to the jth subpopulation on the kth cognitive test in the battery, $b_{0ik}^{(j)}$ and $b_{1ik}^{(j)}$ are the random effects corresponding to the ith patient on the kth test in the cognitive battery if $z_{i\ell}=j$, and $\theta^{(j)}$ is the half-life for the kth cognitive test if $z_{i\ell}=j$. We have chosen to not include additional covariates in the half-life regression equation for simplicity. However, we could include additional covariates such at the number of times the participant has previously completed the kth test in the battery of cognitive tests.

Data were simulated using results from the analysis of the uncensored dataset presented in Table 3.4 in Chapter Three. In Scenario 1 we will define $K=1$, corresponding to only a single test in the given cognitive battery which we set to be the TMTB. In Scenario 2 we will define $K=2$, corresponding to a battery of cognitive tests with two tests consisting of the TMTB and Grooved Pegboard (Dominant Hand). These two tests were chosen due to the magnitude of the posterior estimate of the mean μ. From Table 3.4, the posterior estimate for the mean of the marginal distribution of the TMTB in the cognitively impaired subpopulation was $\widehat{\mu}=-4.08$ and the posterior estimate for the mean of the marginal distribution of the Grooved Pegboard (Dominant Hand) in the cognitively impaired subpopulation was $\widehat{\mu}=-3.99$. Therefore, for the first test in the cognitive battery (TMTB), we will define $\beta_{0,k=1}^{(j=1)}=0$ to correspond to the known baseline mean for the healthy subpopulation and $\beta_{0,k=1}^{(j=2)}=-4.08$ to correspond the estimated mean for the cognitively impaired subpopulation from Table 3.4. For the second test in the cognitive battery (Grooved Pegboard (Dominant Hand)), we will define $\beta_{0,k=2}^{(j=1)}=0$ to correspond

to the known baseline mean for the healthy subpopulation and $\beta_{0,k=2}^{(j=1)} = -3.99$ to correspond to the estimated mean for the cognitively impaired subpopulation from Table 3.4. Then, we know that the standard deviation of the marginal distribution of the TMTB and the Grooved Pegboard (Dominant Hand) are one. Furthermore, the posterior estimate of the standard deviation of the marginal distribution of TMTB in the cognitively impaired subpopulation was $\widehat{\sigma} = 3.71$ and the posterior estimate of the standard deviation of the marginal distribution of Grooved Pegboard (Dominant Hand) in the cognitively impaired subpopulation was $\widehat{\sigma} = 4.39$. However, with the mixed HMM, assuming independence among the fixed effects and error term, the variance at baseline is the sum of the variance of the random intercept and the variance of the error term. Therefore, for the TMTB test for the healthy subpopulation, we defined the marginal distribution of the random intercept to be $\mathcal{N}(0, \sigma = 0.9)$ and the marginal distribution of the error term to be $\mathcal{N}(0, \sigma = \sqrt{1 - 0.9^2})$. Then, for the cognitively impaired subpopulation, we defined the marginal distribution of the random intercept to be $\mathcal{N}(0, \sigma = \sqrt{(3.71^2 - 0.19)})$ and the marginal distribution of the error term to be $\mathcal{N}(0, \sigma = \sqrt{0.19})$. Then, for the Grooved Pegboard (Dominant Hand) test for the healthy subpopulation, we defined the marginal distribution of the random intercept to be $\mathcal{N}(0, \sigma = 0.8)$ and the marginal distribution of the error term to be $\mathcal{N}(0, \sigma = \sqrt{1 - 0.8^2})$. Then, for the cognitively impaired subpopulation, we defined the marginal distribution of the random intercept to be $\mathcal{N}(0, \sqrt{(4.39^2 - 0.6^2)})$ and the marginal distribution of the error term to be $\mathcal{N}(0, \sigma = 0.6)$. Given that the estimated correlation between the TMTB test and the Grooved Pegboard (Dominant Hand) test was near 0, we chose to specify a larger correlation between the two tests and defined the correlation amongst the two tests to be 0.5. Then, for the TMTB test, we defined $\beta_{1,k=1}^{(j=1)} = \beta_{1,k=1}^{(j=2)} = 1$ corresponding to a one standard deviation increase in a patient's cognitive test score if the patient were retested immediately after completion of initial testing. The joint distribution of the random effects for the healthy subpopulation

was given a multivariate normal distribution with mean vector $\mathbf{0} = (0, 0)^\top$, standard deviation vector $\boldsymbol{\sigma}_{\boldsymbol{b}_{i,k=1}^{(j=1)}} = (0.9, 0.5)^\top$ and correlation $\rho_{\boldsymbol{b}_{i,k=1}^{(j=1)}} = -0.7$. For the cognitively impaired subpopulation, the joint distribution of the random effects was given a multivariate normal distribution with mean vector $\mathbf{0} = (0, 0)^\top$, standard deviation vector $\boldsymbol{\sigma}_{\boldsymbol{b}_{i,k=1}^{(j=1)}} = (3.68, 0.5)^\top$ and correlation $\rho_{\boldsymbol{b}_{i,k=1}^{(j=1)}} = -0.7$. Then, for the Grooved Pegboard (Dominant Hand) test, we defined $\beta_{1,k=2}^{(j=1)} = \beta_{1,k=2}^{(j=2)} = 1$ corresponding to a one standard deviation increase in a patient's cognitive test score if the patient were retested immediately after completion of initial testing. The joint distribution of the random effects for the healthy subpopulation was given a multivariate normal distribution with mean vector $\mathbf{0} = (0, 0)^\top$, standard deviation vector $\boldsymbol{\sigma}_{\boldsymbol{b}_{i,k=2}}^{(j)} = (0.8, 0.4)^\top$, and correlation $\rho_{\boldsymbol{b}_{i,k=2}}^{(j)} = -0.5$, $j = 1, 2$. For the cognitively impaired subpopulation, we defined the joint distribution of the random effects as a multivariate normal distribution with mean vector $\mathbf{0} = (0, 0)^\top$, standard deviation vector $\boldsymbol{\sigma}_{\boldsymbol{b}_{i,k=2}}^{(j)} = (4.35, 0.4)^\top$, and correlation $\rho_{\boldsymbol{b}_{i,k=2}}^{(j)} = -0.5$, $j = 1, 2$. We will then assume that the joint distribution of $\boldsymbol{\sigma}_\epsilon^{(j)} = (\sigma_{\epsilon,k=1}^{(j)}, \sigma_{\epsilon,k=2}^{(j)})^\top$ is multivariate normal with mean vector $\mathbf{0} = (0, 0)^\top$, standard deviation vector $\boldsymbol{\sigma}_\epsilon^{(j)} = (\sqrt{0.19}, 0.6)^\top$ and correlation $\rho_\epsilon^{(j)} = 0.5$, $j = 1, 2$. Finally, will let $\theta_k^{(j)} = 0.5$, $j = 1, 2$, $k = 1, 2$ to correspond to a half-life which results in the decay of the practice effect to near zero by 6 months post testing.

Figure 4.2 shows the longitudinal scores on a single cognitive test as well as the respective positions of the average (and 95% confidence limits) cognitive test score for the two latent class subpopulations corresponding to the ith subject.

4.4.2 Statistical Software

All analyses will be performed using Stan in R (version 3.4.3) via the Rstan package (version 2.17.3) (R Core Team, 2016; Stan Development Team, 2016, 2017). The continuous-time mixed HMM will be fit with two subpopulations, as in Chapter Three, corresponding to the cognitively impaired and healthy subpopulations.

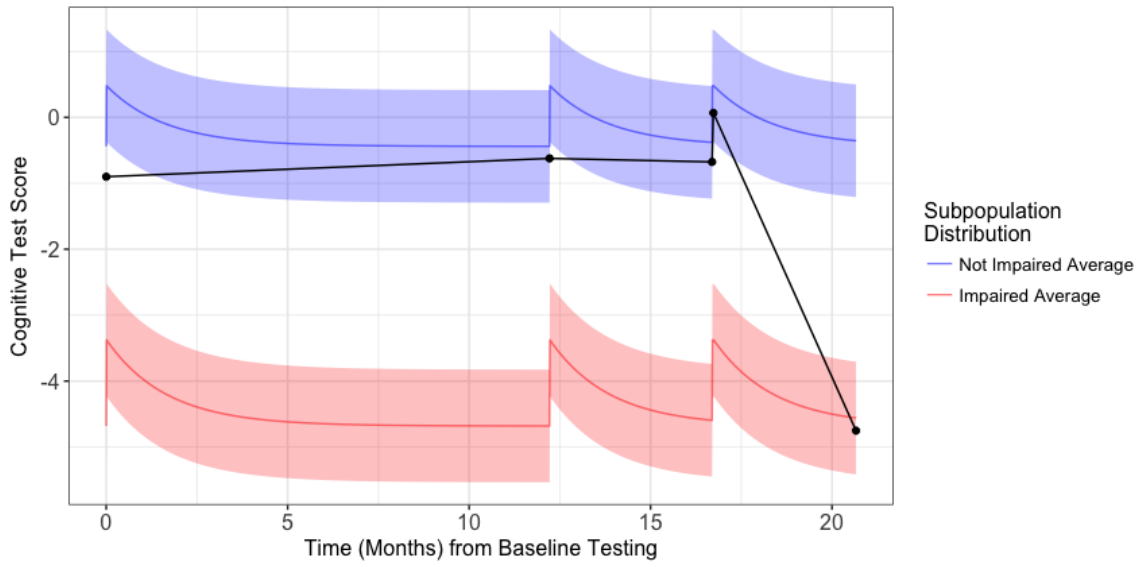


Figure 4.2. Longitudinal cognitive test scores from a randomly selected patient and the corresponding subpopulation distribution with shaded regions denoting the 95% confidence interval for each subpopulation.

Analysis in both scenarios was originally performed for 10,000 iterations with a 5,000 iteration warm-up using three chains in order to ensure that the posterior distribution for all parameters was converging to a stationary distribution. After confirming convergence, the analysis was performed again for 25,000 iterations with a 5,000 iteration burn-in using only a single chain.

4.4.3 Forward and Backwards Algorithm

In order to estimate the parameters in the continuous-time mixed HMM, we must use the forward algorithm as described in section 4.2. The forward algorithm as performed in Stan is defined in Algorithm 1. To estimate the cognitive impairment status of the ith patient at time $t_{i\ell}$, we must use the backwards algorithm as discussed in section 4.2. The backwards algorithm as performed in Stan is defined in Algorithm 2.

Data: $Y_i \in \mathbb{R}^{N\times K}$ and $t_{i\ell}$ for $i = 1,\ldots,N$, $\ell = 1,\ldots,m_i$, $k = 1,\ldots,K$
Result: $\widehat{\alpha}_\ell(j) = P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_{m_i} = y_{m_i}, Z_{i\ell} = j)$

begin
  for $\widehat{\lambda}_j$ at each iteration **do**
    $\widehat{\alpha}_1(z_{ij}) \leftarrow \widehat{\lambda}_{z_j}$
    for $i \in \{1,2,\ldots,N\}$ **do**
      for $\ell \in \{1,2,\ldots,m_i\}$ **do**
        for $j = 1,2$ **do**
          $\widehat{\alpha}_\ell(j) \leftarrow \widehat{\alpha}_\ell(1) f_j(y_{i\ell})$
        end
        for $j = 1,2$ **do**
          $\widehat{\alpha}_\ell(j) \leftarrow \widehat{\alpha}_\ell(j) / \sum_{j=1,2} \widehat{\alpha}_\ell(j)$
        end
        if $\ell < m_i$ **then**
          for $j = 1,2$ **do**
            $\widehat{\alpha}_{\ell+1}(j) \leftarrow \widehat{\alpha}_\ell(j)\widehat{\Gamma}(t_{i(\ell+1)} - t_{i\ell})_{1,j} + \widehat{\alpha}_\ell(2)\widehat{\Gamma}(t_{i(\ell+1)} - t_{i\ell})_{2,j}$
          end
          for $j = 1,2$ **do**
            $\widehat{\alpha}_{\ell+1}(j) \leftarrow \widehat{\alpha}_{\ell+1}(j) / \sum_{j=1,2} \widehat{\alpha}_{\ell+1}(j)$
          end
        end
      end
    end
  end
end

Algorithm 1: Forward Algorithm

4.4.4 Data Model and Priors

The data model in Scenario 1 and Scenario 2 were specified to match the true data models presented in subsection 4.4.1. Using the results from Chapter Three, we used informative priors for appropriate parameters and used weakly informative priors for all other parameters. The data model for both longitudinal test scores was defined to be

$$(Y_{i\ell}|z_{i\ell} = j) \sim \mathcal{N}_k(\mu_{i\ell k}^{(j)}, \sigma_{\epsilon_k}^{(j)}),$$

such that $\mu_{i\ell k}^{(j)} = (\beta_{0k}^{(j)} + b_{0ik}^{(j)})$ at baseline (t_{i1}), and for $t_{i\ell}$, $\ell = 2,\ldots,m_i$ $\mu_{i\ell k}^{(j)} = (\beta_{0k}^{(j)} + b_{0ik}^{(j)}) + (\beta_{1k}^{(j)} + b_{1ik}^{(j)}) \exp\{-(t_{i\ell} - t_{i(\ell-1)}) / \exp[\theta_k^{(j)}]\}$. As was the case in

Data: $Y_i \in \mathbb{R}^{N\times K}$ and t_{ij} for $i = 1, \ldots, N$, $j = 1, \ldots, m_i$, $k = 1, \ldots, K$
Result: $P(Z_{ij} = z_{ij}|Y_1 = y_1, Y_2 = y_2, \ldots, Y_{m_i} = y_{m_i})$

begin
  for $\widehat{\alpha}_\ell$ at each iteration **do**
    $\widehat{\beta}_{m_i}(j) \leftarrow 1$, $j = 1, 2$
    for $i \in \{1, 2, \ldots, N\}$ **do**
      for $\ell \in \{1, 2, \ldots, m_i - 1\}$ **do**
        $u \leftarrow m_i - \ell$
        for $j = 1, 2$ **do**
          $\widehat{\beta}_u(j) \leftarrow \sum_{j=1,2} \widehat{\Gamma}(t_{i(u+1)} - t_{iu})_{1,j} f_j(y_{i,u})\beta_{u+1}(j)$
        end
        for $j = 1, 2$ **do**
          $\widehat{\beta}_u(j) \leftarrow \widehat{\beta}_u(j)/\sum_{j=1,2} \widehat{\beta}_u(j)$
        end
        for $j = 1, 2$ **do**
          $P(Z_{ij} = j|Y_1 = y_1, Y_2 = y_2, \ldots, Y_j = y_j) =$
          $\widehat{\alpha}_u(j)\widehat{\beta}_u(j)/[\widehat{\alpha}_u(j)\widehat{\beta}_u(j)]$
        end
      end
    end
  end
end

Algorithm 2: Backwards Algorithm

Chapter Three, we know that the baseline mean for the healthy subpopulation for each test is zero. Therefore, we will specify the known parameter $\beta_{0,k=1}^{j=1} = \beta_{0,k=2}^{j=1} = 0$. Similarly, we know that the baseline standard deviation of each test is one for the healthy subpopulation. Therefore, we can define $\sigma_k^{(j=1)} = 1$, where $\sigma_k^{(j=1)}$ is the standard deviation of the marginal distribution of the kth test, $k = 1, 2$, in the healthy subpopulation.All prior distributions specified for the parameters in the continuous-time mixed HMM are shown in Figure 4.3. The only additional information we have regarding the values of other parameters in the continuous-time mixed HMM we can obtain from the results in Chapter Three. For the baseline mean of the marginal distribution of the cognitively impaired subpopulation for test $k = 1$ (TMTB), we see from Table 3.4 that the posterior mean is -4.08 with a 95% confidence interval

$$
\begin{aligned}
&\beta_{0,k=1}^{(j=2)} \sim \mathcal{N}(-4.08, 1); \\
&\beta_{0,k=2}^{(j=2)} \sim \mathcal{N}(-3.99, 1); \\
&\beta_{1k}^{(j)} \sim \mathcal{N}(0, 1), && k = 1, 2, \quad j = 1, 2; \\
&\boldsymbol{\Omega}_{\epsilon} \sim \text{LKJ-correlation distribution}(\eta = 1); \\
&\boldsymbol{\epsilon}_i^{(j)} \sim \mathcal{N}\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\text{diag}\left(\boldsymbol{\sigma}_{\epsilon}^{(j)}\right)\right] \boldsymbol{\Omega}_{b_{ik}} \left[\text{diag}\left(\boldsymbol{\sigma}_{\epsilon}^{(j)}\right)\right]\right) && k = 1, 2, \quad j = 1, 2; \\
&\sigma_{\epsilon_k}^{(j=1)} \sim \text{Uniform}(0, 1), && k = 1, 2; \\
&\sigma_{\epsilon_k}^{(j=2)} \sim \text{Lognormal}(0, 1), && k = 1, 2; \\
&\boldsymbol{b}_{ik}^{(j)} \sim \mathcal{N}\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\text{diag}\left(\boldsymbol{\sigma}_{b_{ik}}^{(j)}\right)\right] \boldsymbol{\Omega}_{b_{ik}} \left[\text{diag}\left(\boldsymbol{\sigma}_{b_{ik}}^{(j)}\right)\right]\right); \\
&\sigma_{b_{0i,k}}^{(j=1)} = \sqrt{1 - \left(\sigma_{\epsilon_k}^{(j=1)}\right)^2}, && k = 1, 2; \\
&\sigma_{b_{0i,k}}^{(j=2)} = \sqrt{\left(\sigma_k^{(j=2)}\right)^2 - \left(\sigma_{\epsilon_k}^{(j=2)}\right)^2}, && k = 1, 2; \\
&\sigma_{b_{1i,k}}^{(j=1)} \sim \text{Lognormal}(0, 1), && k = 1, 2; \\
&\sigma_{k=1}^{(j=2)} \sim \text{Lognormal}(1.3, 0.5); \\
&\sigma_{k=2}^{(j=2)} \sim \text{Lognormal}(1.47, 0.5); \\
&\sigma_{b_{1i,k}}^{(j=2)} \sim \text{Lognormal}(0, 1), && k = 1, 2; \\
&\boldsymbol{\Omega}_{b_{ik}} \sim \text{LKJ-correlation distribution}(\eta = 1); \\
&q_j \sim \text{Half-Normal}(0, 1), && j = 1, 2; \\
&\theta_k^{(j)} \sim \mathcal{N}(0, 1) && k = 1, 2, \quad j = 1, 2; \\
&\boldsymbol{\lambda} \sim \text{Dirichlet}(24, 21);
\end{aligned}
$$

Figure 4.3. Prior distributions for all parameters in the continuous-time mixed hidden Markov model

between $(-5.72, -2.47)$. Therefore, we will define an informative prior for the baseline mean of the marginal distribution of the cognitive impaired subpopulation for the TMTB test in the form of a $\mathcal{N}(-4.08, 1)$ in accordance with the posterior estimates in Table 3.4. Similarly, for the baseline mean of the marginal distribution of the cognitively impaired subpopulation for test $k = 2$ (Grooved Pegboard (Dominant Hand)), we see from Table 3.4 that the posterior mean is -3.99 with a 95% confidence interval between $(-5.87, -2.13)$. Therefore, we will define an informative prior for the baseline mean of the marginal distribution of the cognitive impaired subpopulation for the Grooved Pegboard (Dominant Hand) test in the form of a $\mathcal{N}(-3.99, 1)$. Additionally, the estimated posterior mean for $\sigma_{k=1}^{j=2}$ in Table 3.4 was 3.71 with a 95%

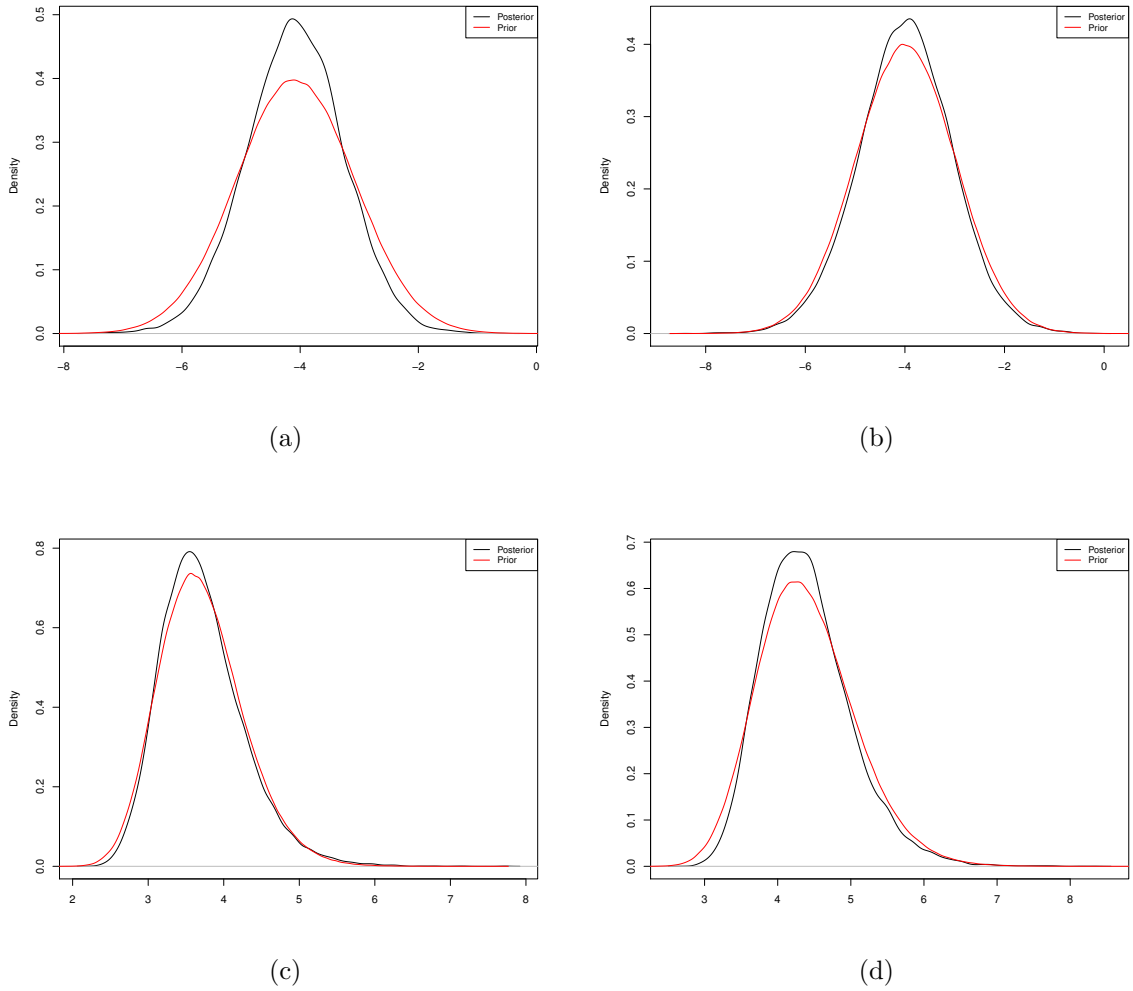


Figure 4.4. Posterior distribution obtained in Chapter Three and selected prior for the mean of the marginal distribution for the (a) TMTB test and (b) Grooved Pegboard (Dominant Hand) test and the standard deviation of the marginal distribution for the (c) TMTB test and (d) Grooved Pegboard (Dominant Hand) test.

credible interval of (2.82, 4.97). The similarity of the prior distributions discussed for the mean of the marginal distribution at baseline for the cognitively impaired subpopulation for the TMTB test and the Grooved Pegboard (Dominant Hand) test as well as the standard deviation of the marginal distribution at baseline for the cognitively impaired subpopulation for the TMTB test and the Grooved Pegboard (Dominant Hand) test used in our analysis and the estimated posterior distribution obtained in Chapter Three are shown in Figure 4.4. In order to provide a informative prior based on the estimated posterior distribution, we specified a lognormal(1.3, 0.5).

Similarly, the estimated posterior mean for $\sigma_{k=2}^{j=2}$ in Table 3.4 was 4.39 with a 95% credible interval of (3.40, 5.78). In order to provide a informative prior based on the estimated posterior distribution, we specified a lognormal(1.47, 0.5). Lastly, we specified an informative prior distribution for $\boldsymbol{\lambda}$ in the form of a Dirichlet(21, 24) based on the number of patients estimated to belong to each subpopulation in Chapter 4. For all other parameters, weakly informative priors were chosen to reflect the limited information regarding potential parameter values.

4.5 Results

The parameter estimates and 95% credible intervals for both Scenario 1 and Scenario 2 are shown in Table 4.1. Based on the results presented, for both Scenario 1 and Scenario 2 we see that the true value of all parameters included in the continuous-time HMM are contained within their respective 95% credible interval. However, we see that in Scenario 1 that the 95% credible interval for $\sigma_{b_{1,k=1}}^{j=1}$ is very wide. Similarly, in Scenario 2 the 95% credible interval for $\sigma_{b_{1,k=1}}^{j=1}$ and $\sigma_{b_{1,k=1}}^{j=1}$ have a large range. Lastly, we see that $\rho_{b,k=1}$ in Scenario 1 and $\rho_{b,k=2}$ in Scenario 2 include positive value, but their respective posterior means are indeed negative. Once real data are ready to be analyzed, if we can obtain illicit more informative priors from an expert, we may be able to improve the accuracy of posterior estimates.

Additionally, we can examine the predictive accuracy of the model to correctly infer the cognitive status of patients at each time point t_{ij}. For Scenario 1, we find that out of a total of $n \times m_i = 500$ total observed instances at which a patient's cognitive status is inferred using the Bayesian continuous-time HMM, the cognitive status is incorrectly inferred 69 times (13.8%). For Scenario 2, 19 (3.8%) patients cognitive status is incorrectly inferred.

Table 4.1. Parameters in continuous-time mixed hidden Markov model and the respective posterior estimates for Scenario 1 ($K = 1$) and Scenario 2 ($K = 2$); j refers to the subpopulation, $j = 1, 2$ and k refers to the cognitive test number in the given battery $k = 1, \ldots, K$.

				Scenario 1		Scenario 2	
Parameter	j	k	True Value	Posterior Mean	95% Credible Interval	Posterior Mean	95% Credible Interval
β_0	2	1	−4.08	−4.16	(−4.87, −3.48)	−3.51	(−4.16, −2.86)
β_1	1	1	1	1.44	(0.45, 2.71)	0.99	(0.50, 1.64)
β_1	2	1	1	1.14	(0.71, 1.60)	0.92	(0.65, 1.24)
β_0	2	2	−3.99			−3.70	(−4.54, −2.87)
β_1	1	2	1			0.94	(0.34, 1.67)
β_1	2	2	1			1.14	(0.74, 1.60)
σ_{b_0}	1	1	0.9	0.87	(0.82, 0.91)	0.92	(0.90, 0.95)
σ_{b_0}	2	1	3.71	3.79	(3.34, 4.31)	3.53	(3.11, 3.99)
σ_{b_1}	1	1	0.5	0.65	(0.10, 1.82)	0.46	(0.11, 1.05)
σ_{b_1}	2	1	0.5	0.60	(0.15, 1.11)	0.28	(0.09, 0.54)
σ_{b_0}	1	2	0.8			0.73	(0.61, 0.82)
σ_{b_0}	2	2	4.39			4.63	(4.09, 5.23)
σ_{b_1}	1	2	0.4			0.64	(0.11, 1.57)
σ_{b_1}	2	2	0.4			0.55	(0.13, 1.11)
ρ_b	1, 2	1	−0.7	−0.31	(−0.87, 0.32)	−0.70	(−1.00, −0.13)
ρ_b	1, 2	2	−0.5			−0.40	(−0.87, 0.27)
σ_ϵ	1	1	0.44	0.48	(0.41, 0.57)	0.38	(0.33, 0.44)
σ_ϵ	2	1	0.44	0.42	(0.37, 0.48)	0.43	(0.39, 0.47)
σ_ϵ	1	2	0.6			0.68	(0.58, 0.80)
σ_ϵ	2	2	0.6			0.60	(0.54, 0.66)
ρ_ϵ	1, 2	1,2	0.5			0.46	(0.37, 0.55)
θ	1	1	0.5	−0.43	(−1.35, 0.59)	0.28	(−0.44, 1.09)
θ	1	2	0.5			0.43	(−0.61, 1.30)
θ	2	1	0.5	0.63	(0.10, 1.15)	0.39	(−0.21, 1.00)
θ	2	2	0.5			0.33	(−0.17, 0.82)
q_1	-	-	0.6	0.54	(0.27, 1.07)	0.96	(0.50, 1.84)
q_2	-	-	0.4	0.33	(0.17, 0.66)	0.50	(0.26, 0.96)
λ_1	-	-	0.5	0.53	(0.45, 0.62)	0.47	(0.39, 0.56)

4.6 Concluding Remarks

As cognitive impairment status over time becomes a quantity of interest in clinical research and clinical trials, a method for accurately inferring cognitive impairment based on serial administrations of a battery of cognitive tests is crucial. Moreover, such a method must also be able to account for the presence of a practice effect caused by repeated administration of a given cognitive test. We have proposed the use of a Bayesian continuous-time mixed hidden Markov model to infer cognitive impairment from a battery of cognitive tests administered serially. Additionally, we have incorporated an additional term in the distribution of the battery of cognitive test scores conditional on the latent subpopulation adapted from the half-life regression equation proposed by Settles and Meeder (2016) for recall which is able to capture the presence of a practice effect and can incorporate various covariates which may affect the magnitude and/or rate of decay of the practice effect.

Due to the unavailability of longitudinal scores from the battery of cognitive tests administered to pediatric MS patients treated at the Pediatric Demyelinating Diseases Clinic of Children's Health Dallas, we were forced to simulate longitudinal data based on a subset of the results presented in Chapter Three. We first simulated longitudinal cognitive test scores for a single cognitive test from a mixed effects model containing the additional practice effect term and found that the 95% credible intervals of the posterior estimates contained the true value of all parameters contained in the model. Additionally, the cognitive status was inferred incorrectly for only 0.65% of the total instances at which a cognitive test was administered. Then, we simulated longitudinal cognitive test scores for a battery of two cognitive tests from a multivariate mixed effects model containing the additional practice effect term and found that all except for three of the 95% credible intervals of the posterior estimates contained the true value of their respective parameter. However, more importantly, the cognitive status was inferred correctly for every instance at which a cognitive test

was administered. Therefore, the Bayesian continuous-time mixed hidden Markov model performed well for the simulated data.

The main limitation of the current study is the absence of actual longitudinal scores from a battery of cognitive tests administered to a sample of the patient population of interest. Preliminary data is crucial to the development of an appropriate clinical trial. Additionally, the proposed method was investigated in a battery of cognitive tests much smaller than the true battery of cognitive tests used for pediatric MS patients treated at the Pediatric Demyelinating Diseases Clinic of Children's Health Dallas due to time constraints. As the number of tests in a given battery increases, the computational complexity of the model increases with limited, if any, information available to provide more informative priors for any of the parameters. However, the proposed model may be simplified by limiting the number of random effects, but this leads to the need for estimating the correlation matrix for the errors terms for each cognitive test.

Future work will focus on obtaining and analyzing actual longitudinal results from the battery of cognitive tests presented in Chapter Three, an effort that has already begun. After which, we will construct a simulation study similar to that performed in Chapter Three to examine the predictive accuracy of the proposed method. After which, we will assist in the development of a clinical study investigating the effect of therapeutic agents on patients' cognitive impairment status over time.

CHAPTER FIVE

Conclusion

The understanding of cognitive impairment in patients suffering from various diseases and its impact on the patient's and their family's quality-of-life has led to an increased interest in researching this symptom of the disease. Longitudinal studies are currently underway investigating cognitive impairment in pediatric Multiple Sclerosis, as well as early discussions related to clinical trials investigating the impact of therapeutic intervention on cognitive impairment in this patient population. However, the diagnosis of the latent class status of cognitive impairment has been flawed and must be resolved before proper analysis of longitudinal cognitive testing data and the design of longitudinal clinical trials.

In Chapter Three we investigated the use of a Bayesian multivariate finite mixture model to infer cognitive impairment based on a battery of cognitive tests administered at baseline. The proposed method and currently used methods for inferring cognitive impairment based on a battery of cognitive tests were applied to baseline results from pediatric Multiple Sclerosis patients seen at the Pediatric Demyelinating Diseases Clinic of Children's Health Dallas. In order to examine the sensitivity and specificity of the proposed method and the currently used methods, a simulation study was constructed to reflect the results from the pediatric MS patient data. In the simulation study we find that the Bayesian multivariate mixture model has a greater sensitivity and specificity relative to the currently used methods. Additionally, we examined the effect of censored cognitive test scores on parameter estimates and the inferred cognitive impairment status. We found that despite accounting for censoring, the resulting parameter estimates were biased when analyzing censored cognitive test scores relative to the analysis of uncensored cognitive test scores.

Finally, in Chapter Four we propose the use of a Bayesian continuous-time mixed hidden Markov model accounting for the practice effect to infer cognitive impairment status in serial administrations of a battery of cognitive tests. We simulated data for a serial administration of a single cognitive test and a battery of two cognitive tests based on a subset of the results obtained in Chapter Three to examine the ability of the Bayesian continuous-time mixed hidden Markov model. The Bayesian continuous-time mixed hidden Markov model correctly inferred cognitive impairment for 99.35% of administration of the single cognitive test and 100% of administrations of the battery of two cognitive tests in the presence of a practice effect.

Given the predictive ability of the proposed methods to accurately infer cognitive impairment, we are able to move forward with longitudinal studies of cognitive impairment in the pediatric Multiple Sclerosis patient population. As sample sizes amass to greater numbers, we can apply the proposed methods and obtain better estimates of the correlation amongst cognitive tests in the given battery, the transition probabilities, and the impact of the practice effect. After which, we can accurately estimate the necessary sample size to conduct a longitudinal clinical trial investigating the impact of therapeutic intervention. Lastly, we can then investigate the association of longitudinal biomarker measurements on cognitive impairment status and use such measurements to make predictions that a given patient will become cognitively impaired at a future time point, a task that can be accomplished using the methods investigated in Chapter Two. Based on the results presented in Chapter Two, it is clear that the joint modeling approach and landmark approach are most appropriate with the choice amongst the two methods depending on biologic variability. Additional research will be needed to apply the dynamic prediction methods to the scenario where patients are able to switch back-and-forth between the two latent class status. Accomplishing such task will dramatically improve the understanding and treatment of cognitive impairment in pediatric Multiple Sclerosis patients result-

ing in an alleviation of this symptom of disease and lead to an improvement in the quality-of-life of all those impacted by this disease.

In addition to the ability of the proposed methods to accurately infer cognitive impairment based on the current battery of cognitive tests in pediatric Multiple Sclerosis patients, the proposed methods are suitable for any battery of cognitive tests thought to capture cognitive impairment. Furthermore, the proposed methods can be implemented to infer cognitive impairment in other conditions which affect cognition such as Alzheimer's disease and Parkinson's disease. The Bayesian models created for the analysis of pediatric Multiple Sclerosis data can be easily adapted to any battery of cognitive tests for any condition. Our efforts have led to an improved and statistically valid method for inferring cognitive impairment from a battery of cognitive tests with broad applicability in the research and treatment of cognitive impairment.

APPENDIX

APPENDIX A

Posterior Estimates of the Correlation Matrix in the Analysis of Censored and Uncensored Data

A.1 Posterior Estimates of the Correlation Matrix

In Chapter 3, we defined the multivariate distribution of the scores obtained from the battery of cognitive tests conditional on the subpopulation membership, z_i, to be

$$f(\boldsymbol{y}|\boldsymbol{\theta}_{z_i}) = \frac{1}{(2\pi)^{K/2}|(\text{diag}(\boldsymbol{\sigma}_{z_i}))\boldsymbol{\Omega}(\text{diag}(\boldsymbol{\sigma}_{z_i}))|^{1/2}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_{z_i})^{\mathrm{T}}\left[(\text{diag}(\boldsymbol{\sigma}_{z_i}))\boldsymbol{\Omega}(\text{diag}(\boldsymbol{\sigma}_{z_i}))\right]^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_{z_i})\right\}.$$

The resulting posterior estimates for all other parameters in the finite multivariate mixture model have been presented in Chapter 3, with the exception of $\boldsymbol{\Omega}$ due to its size. We present the posterior estimates for the off-diagonal elements of $\boldsymbol{\Omega}$ from the analysis of the censored and uncensored dataset in Tables A.1-A.2 and Tables A.3-A.4, respectively.

Let $\boldsymbol{\Omega}^*$ denote the upper off-diagonal portion of $\boldsymbol{\Omega}$. Furthermore, let $\boldsymbol{\Omega}_1^*$ denote the first seven columns of $\boldsymbol{\Omega}^*$ and let $\boldsymbol{\Omega}_2^*$ denote the last seven columns of $\boldsymbol{\Omega}^*$, such that $\boldsymbol{\Omega}^* = [\boldsymbol{\Omega}_1^* \vdots \boldsymbol{\Omega}_2^*]$. Table A.1 presents the posterior estimates for $\boldsymbol{\Omega}_1^*$ from the censored dataset, table A.2 presents the posterior estimates for $\boldsymbol{\Omega}_2^*$ from the censored dataset, table A.3 presents the posterior estimates for $\boldsymbol{\Omega}_1^*$ for the uncensored dataset, and table A.4 presents the posterior estimates for $\boldsymbol{\Omega}_2^*$ for the uncensored dataset.

Table A.1. Posterior mean and 95% credible interval of upper triangular, non unity elements of the first partition of the correlation matrix, $\mathbf{\Omega}_1^*$, estimated for the censored dataset, such that $\mathbf{\Omega}^* = [\mathbf{\Omega}_1^* \vdots \mathbf{\Omega}_2^*]$.

Cognitive Test	CVLT Trial 1	CVLT Trial 5	CVLT Long Delay	CVLT Rec. Hits	Grooved Pegboard (Dominant)	Grooved Pegboard (Non-Dominant)	Digit Span
CVLT Total	0.60 (0.42, 0.74)	0.52 (0.33, 0.68)	0.62 (0.41, 0.77)	0.30 (0.03, 0.52)	0.16 (−0.14, 0.40)	0.08 (−0.20, 0.35)	0.29 (0.04, 0.51)
CVLT Trial 1		0.18 (−0.05, 0.39)	0.23 (−0.02, 0.44)	0.06 (−0.19, 0.30)	−0.07 (−0.35, 0.19)	0.06 (−0.22, 0.31)	0.24 (0.00, 0.45)
CVLT Trial 5			0.59 (0.38, 0.74)	0.33 (0.09, 0.54)	0.03 (−0.20, 0.26)	0.23 (−0.04, 0.47)	0.36 (0.13, 0.55)
CVLT Long Delay				0.33 (0.06, 0.57)	0.29 (0.00, 0.53)	0.24 (−0.08, 0.52)	0.31 (0.05, 0.53)
CVLT Rec. Hits					0.06 (−0.22, 0.32)	0.13 (−0.19, 0.43)	0.25 (−0.02, 0.49)
Grooved Pegboard (Dominant)						0.36 (0.08, 0.59)	−0.08 (−0.32, 0.17)
Grooved Pegboard (Non-Dominant)							0.21 (−0.07, 0.46)

Table A.2. Posterior mean and 95% credible interval of upper triangular, non unity elements of the second partition of the correlation matrix, $\boldsymbol{\Omega}_2^*$, estimated for the censored dataset, such that $\boldsymbol{\Omega}^* = [\boldsymbol{\Omega}_1^* \vdots \boldsymbol{\Omega}_2^*]$.

Cognitive Test	Beery VMI	Beery VP	TMTA	TMTB	Symbol Search	SDMT	Letter Fluency
CVLT Total	0.04 (−0.19, 0.26)	0.23 (−0.05, 0.48)	0.05 (−0.21, 0.30)	0.24 (−0.05, 0.47)	0.18 (−0.09, 0.42)	0.31 (0.07, 0.52)	0.26 (0.01, 0.48)
CVLT Trial 1	0.12 (−0.10, 0.33)	0.09 (−0.18, 0.35)	0.19 (−0.07, 0.42)	0.18 (−0.09, 0.42)	−0.02 (−0.28, 0.23)	0.33 (0.09, 0.53)	0.08 (−0.17, 0.31)
CVLT Trial 5	0.60 (0.42, 0.74)	0.60 (0.42, 0.74)	0.60 (0.42, 0.74)	0.60 (0.42, 0.74)	0.60 (0.42, 0.74)	0.60 (0.42, 0.74)	0.60 (0.42, 0.74)
CVLT Long Delay	0.19 (−0.04, 0.39)	0.23 (−0.03, 0.46)	−0.03 (−0.28, 0.21)	0.07 (−0.17, 0.30)	0.04 (−0.22, 0.28)	0.08 (−0.15, 0.30)	0.06 (−0.18, 0.29)
CVLT Rec. Hits	0.06 (−0.18, 0.28)	0.11 (−0.19, 0.40)	−0.01 (−0.27, 0.26)	0.19 (−0.12, 0.45)	0.20 (−0.09, 0.46)	0.30 (0.06, 0.52)	0.20 (−0.07, 0.44)
Grooved Pegboard (Dominant)	0.08 (−0.17, 0.31)	0.11 (−0.21, 0.41)	−0.11 (−0.39, 0.18)	−0.04 (−0.31, 0.23)	0.21 (−0.09, 0.47)	0.16 (−0.10, 0.40)	0.13 (−0.15, 0.39)
Grooved Pegboard (Non-Dominant)	0.16 (−0.07, 0.38)	0.22 (−0.13, 0.54)	0.09 (−0.21, 0.38)	−0.06 (−0.33, 0.23)	0.20 (−0.10, 0.47)	0.08 (−0.21, 0.34)	0.11 (−0.19, 0.39)
Digit Span	0.04 (−0.18, 0.26)	0.02 (−0.28, 0.30)	0.00 (−0.25, 0.26)	0.25 (0.00, 0.48)	0.35 (0.09, 0.56)	0.13 (−0.11, 0.36)	0.26 (0.00, 0.48)
Beery VMI		0.27 (0.02, 0.48)	0.01 (−0.22, 0.24)	0.20 (−0.02, 0.41)	0.02 (−0.21, 0.26)	0.07 (−0.15, 0.28)	0.01 (−0.21, 0.24)
Beery VP			0.25 (−0.07, 0.52)	0.12 (−0.17, 0.40)	0.16 (−0.18, 0.45)	0.05 (−0.24, 0.33)	−0.01 (−0.31, 0.29)
TMTA				0.46 (0.21, 0.67)	0.25 (−0.04, 0.52)	0.08 (−0.17, 0.34)	0.05 (−0.22, 0.32)
TMTB					0.31 (0.04, 0.55)	0.17 (−0.07, 0.41)	0.14 (−0.14, 0.41)
Symbol Search						0.31 (0.05, 0.54)	0.08 (−0.20, 0.35)
SDMT							0.14 (−0.12, 0.38)

Table A.3. Posterior mean and 95% credible interval of upper triangular, non unity elements of the first partition of the correlation matrix, $\mathbf{\Omega}_1^*$, estimated for the uncensored dataset, such that $\mathbf{\Omega}^* = [\mathbf{\Omega}_1^* \vdots \mathbf{\Omega}_2^*]$.

Cognitive Test	CVLT Trial 1	CVLT Trial 5	CVLT Long Delay	CVLT Rec. Hits	Grooved Pegboard (Dominant)	Grooved Pegboard (Non-Dominant)	Digit Span
CVLT Total	0.62 (0.44, 0.75)	0.52 (0.32, 0.68)	0.61 (0.41, 0.76)	0.27 (0.00, 0.51)	0.16 (−0.07, 0.38)	0.11 (−0.14, 0.35)	0.27 (0.02, 0.49)
CVLT Trial 1		0.18 (−0.04, 0.38)	0.23 (0.00, 0.45)	0.05 (−0.20, 0.30)	−0.03 (−0.23, 0.18)	0.08 (−0.16, 0.31)	0.23 (−0.01, 0.45)
CVLT Trial 5			0.60 (0.41, 0.74)	0.32 (0.06, 0.54)	0.03 (−0.19, 0.25)	0.18 (−0.06, 0.41)	0.36 (0.13, 0.56)
CVLT Long Delay				0.32 (0.04, 0.56)	0.22 (−0.02, 0.45)	0.15 (−0.12, 0.40)	0.30 (0.04, 0.52)
CVLT Rec. Hits					0.09 (−0.15, 0.32)	0.09 (−0.19, 0.36)	0.24 (−0.05, 0.49)
Grooved Pegboard (Dominant)						0.55 (0.35, 0.70)	−0.02 (−0.24, 0.21)
Grooved Pegboard (Non-Dominant)							0.12 (−0.14, 0.37)

Table A.4. Posterior mean and 95% credible interval of upper triangular, non unity elements of the second partition of the correlation matrix, $\mathbf{\Omega}_2^*$, estimated for the uncensored dataset, such that $\mathbf{\Omega}^* = [\mathbf{\Omega}_1^* \vdots \mathbf{\Omega}_2^*]$.

Cognitive Test	Beery VMI	Beery VP	TMTA	TMTB	Symbol Search	SDMT	Letter Fluency
CVLT Total	0.06 (−0.17, 0.27)	0.22 (−0.05, 0.46)	−0.02 (−0.27, 0.24)	0.30 (0.06, 0.50)	0.19 (−0.07, 0.42)	0.30 (0.06, 0.51)	0.25 (0.01, 0.47)
CVLT Trial 1	0.13 (−0.09, 0.34)	0.12 (−0.15, 0.37)	0.19 (−0.07, 0.41)	0.21 (−0.01, 0.42)	0.00 (−0.24, 0.24)	0.32 (0.09, 0.53)	0.09 (−0.15, 0.31)
CVLT Trial 5	0.17 (−0.05, 0.38)	0.21 (−0.05, 0.45)	−0.08 (−0.32, 0.16)	0.11 (−0.12, 0.33)	0.05 (−0.20, 0.29)	0.06 (−0.17, 0.29)	0.04 (−0.21, 0.28)
CVLT Long Delay	0.08 (−0.16, 0.30)	0.07 (−0.21, 0.35)	−0.04 (−0.30, 0.23)	0.24 (0.00, 0.46)	0.20 (−0.07, 0.45)	0.31 (0.06, 0.52)	0.19 (−0.06, 0.42)
CVLT Rec. Hits	0.06 (−0.19, 0.30)	0.10 (−0.23, 0.41)	−0.18 (−0.45, 0.12)	0.07 (−0.18, 0.32)	0.18 (−0.12, 0.45)	0.14 (−0.13, 0.40)	0.12 (−0.16, 0.39)
Grooved Pegboard (Dominant)	0.12 (−0.09, 0.32)	0.24 (−0.03, 0.48)	0.06 (−0.17, 0.28)	0.18 (−0.04, 0.38)	0.17 (−0.07, 0.39)	0.14 (−0.08, 0.35)	0.03 (−0.22, 0.28)
Grooved Pegboard (Non-Dominant)	0.08 (−0.15, 0.29)	0.29 (−0.03, 0.58)	0.02 (−0.24, 0.28)	0.15 (−0.09, 0.36)	0.16 (−0.11, 0.41)	0.02 (−0.24, 0.26)	0.08 (−0.18, 0.33)
Digit Span	0.03 (−0.20, 0.25)	0.00 (−0.28, 0.28)	−0.01 (−0.27, 0.25)	0.30 (0.07, 0.50)	0.35 (0.09, 0.56)	0.12 (−0.13, 0.46)	0.26 (0.02, 0.49)
Beery VMI		0.23 (−0.03, 0.46)	−0.04 (−0.27, 0.19)	0.19 (−0.02, 0.39)	0.01 (−0.22, 0.25)	0.06 (−0.17, 0.28)	−0.01 (−0.24, 0.23)
Beery VP			0.23 (−0.09, 0.50)	0.24 (−0.04, 0.49)	0.14 (−0.19, 0.42)	0.05 (−0.23, 0.33)	0.01 (−0.29, 0.29)
TMTA				0.36 (0.14, 0.56)	0.23 (−0.04, 0.49)	0.02 (−0.23, 0.28)	0.04 (−0.22, 0.30)
TMTB					0.41 (0.16, 0.61)	0.25 (0.02, 0.46)	0.07 (−0.19, 0.31)
Symbol Search						0.29 (0.04, 0.52)	0.10 (−0.18, 0.35)
SDMT							0.14 (−0.11, 0.38)

BIBLIOGRAPHY

Amato, M., Goretti, B., Ghezzi, A., Lori, S., Zipoli, V., Portaccio, E., Moiola, L., Falautano, M., De Caro, M., Lopez, M., et al. (2008). Cognitive and psychosocial features of childhood and juvenile ms. *Neurology*, 70(20):1891–1897.

Beglinger, L. J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D. A., Crawford, J., Fastenau, P. S., and Siemers, E. R. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Archives of Clinical Neuropsychology*, 20(4):517–529.

Berthelson, L., Mulchan, S. S., Odland, A. P., Miller, L. J., and Mittenberg, W. (2013). False positive diagnosis of malingering due to the use of multiple effort tests. *Brain Injury*, 27(7-8):909–916.

Beyersmann, J., Allignol, A., and Schumacher, M. (2011). *Competing risks and multistate models with R*. Springer Science & Business Media.

Calamia, M., Markon, K., and Tranel, D. (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, 26(4):543–570.

Collett, D. (2015). *Modelling survival data in medical research*. CRC press.

Collie, A., Maruff, P., Darby, D. G., and McSTEPHEN, M. (2003). The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test–retest intervals. *Journal of the International Neuropsychological Society*, 9(3):419–428.

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

Cox, D. R. (1992). Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer.

Crawford, J. R., Garthwaite, P. H., and Gault, C. B. (2007). Estimating the percentage of the population with abnormally low scores (or abnormally large score differences) on standardized neuropsychological test batteries: A generic method with applications. *Neuropsychology*, 21(4):419.

Dikmen, S. S., Heaton, R. K., Grant, I., and Temkin, N. R. (1999). Test–retest reliability and practice effects of expanded halstead–reitan neuropsychological test battery. *Journal of the International Neuropsychological Society*, 5(4):346–356.

Ebbinghaus, H. (2013). Memory: A contribution to experimental psychology. *Annals of neurosciences*, 20(4):155.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL.

Huang, X., Yan, F., Ning, J., Feng, Z., Choi, S., and Cortes, J. (2016). A two-stage approach for dynamic prediction of time-to-event distributions. *Statistics in medicine*.

Huizenga, H. M., Agelink van Rentergem, J. A., Grasman, R. P., Muslimovic, D., and Schmand, B. (2016). Normative comparisons for large neuropsychological test batteries: user-friendly and sensitive solutions to minimize familywise false positives. *Journal of clinical and experimental neuropsychology*, 38(6):611–629.

Huizenga, H. M., Smeding, H., Grasman, R. P., and Schmand, B. (2007). Multivariate normative comparisons. *Neuropsychologia*, 45(11):2534–2542.

Ingraham, L. J. and Aiken, C. B. (1996). An empirical approach to determining criteria for abnormality in test batteries with multiple measures. *Neuropsychology*, 10(1):120.

Jackson, C. H. et al. (2011). Multi-state models for panel data: the msm package for r. *Journal of Statistical Software*, 38(8):1–29.

Julian, L., Serafin, D., Charvet, L., Ackerson, J., Benedict, R., Braaten, E., Brown, T., O'Donnell, E., Parrish, J., Preston, T., et al. (2013). Cognitive impairment occurs in children and adolescents with multiple sclerosis: results from a united states network. *Journal of child neurology*, 28(1):102–107.

Lee, G. and Scott, C. (2012). Em algorithms for multivariate gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, 56(9):2816–2829.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001.

MacAllister, W., Belman, A., Milazzo, M., Weisbrot, D., Christodoulou, C., Scherl, W., Preston, T., Cianciulli, C., and Krupp, L. (2005). Cognitive functioning in children and adolescents with multiple sclerosis. *Neurology*, 64(8):1422–1425.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1–33.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.

Settles, B. and Meeder, B. (2016). A trainable spaced repetition model for language learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1848–1858.

Stan Development Team (2016). RStan: the R interface to Stan. R package version 2.14.1.

Stan Development Team (2017). Stan modeling language users guide and reference manual. Version 2.16.0.

Tan, A., Hague, C., Greenberg, B. M., and Harder, L. (2017). Neuropsychological outcomes of pediatric demyelinating diseases: a review. *Child Neuropsychology*, pages 1–23.

Tsiatis, A. A. and Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88(2):447–458.

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(1):70–85.

Van Montfort, K., Oud, J. H., and Satorra, A. (2010). *Longitudinal research with latent variables*. Springer Science & Business Media.

Westfall, P. H. and Young, S. (2002). Resampling-based multiple testing.

Yeh, E. A., Chitnis, T., Krupp, L., Ness, J., Chabas, D., Kuntz, N., and Waubant, E. (2009). Pediatric multiple sclerosis. *Nature Reviews Neurology*, 5(11):621–631.

Zucchini, W., MacDonald, I. L., and Langrock, R. (2016). *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC.