ESTIMATING THE CONDITIONAL HAZARD FUNCTION OF JOINT LATENT CLASS MIXED MODELS USING HAZARD REGRESSION WITH APPLICATIONS TO PSYCHOLOGY AND NEUROSCIENCE

Joshua P. Rutsohn

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2023

Approved by:

Young Truong

Jessica Girault

Rebecca Grzadzinski

Feng-Chang Lin

Daniela Sotres-Alvarez

ABSTRACT

Joshua P. Rutsohn: Estimating the Conditional Hazard Function of Joint Latent Class Mixed Models using Hazard Regression with Applications to Psychology and Neuroscience
(Under the direction of Young Truong)

Research within the fields psychology and neuroscience often have interest in estimating the change of a latent variable over repeated measurements. While the inference of interest lies within this latent variable, observed variables that express this latent variable are instead measured. If the latent variable is observed by a change in a specific biomarker as well as the time-to-event, then a joint model may be more suitable for estimating the change of the latent variable than a typical single-model approach. Given a heterogeneous sample comprising a mixture of classes of this latent variable, a joint latent class mixed model may prove effective.

Within this joint model, the Cox proportional hazards model is a popular choice for modeling the time-to-event due to its robustness and minimal assumptions. However, the Cox model still requires a proportionality assumption. Hazard regression (HARE) was developed with relaxing this assumption in mind. HARE uses an adaptive B-spline method to estimate the conditional log-hazard function of the survival model not requiring that the hazard function follow this proportionality assumption. The B-splines in HARE can take the form of covariates of interest, time, or a tensor product of the two. An adaptive regression method estimates the conditional log-hazard function via a partial likelihood method.

The purpose of this proposal is to introduce the HARE methodology to estimating the class-specific conditional log-hazard function of a joint latent class mixed model with applications to psychology and neuroscience research. This method is then tested on a study of simulated data, on a subset of data from the Paquid longitudinal cohort study, data from an experiment assessing visual information from infants as they engage in a social arousal task, and longitudinal data

from the Infant Brain Imaging Study. The novelty and utility of HARE within the joint latent class mixed model are explored and discussed.

To my father, Philip, for always being a perfect role model both personal and professional. To my brother, Matt, for always being my best friend.

ACKNOWLEDGEMENTS

Unfortunately, I cannot thank everyone who deserves acknowledgement in this process because there is not nearly enough space to do so. I want to begin by thanking my friend and mentor, Dr. Young Truong, who has been my academic and dissertation advisor throughout my doctoral education. I appreciate all of your help from day one, and I look forward to continue working with you in the future. I want to thank my other committee members, Dr. Jessica Girault, Dr. Rebecca Grzadzinski, Dr. Feng-Chang Lin, and Dr. Daniela Sotres-Alvarez. Whether it be through education, collaboration, or both, your guidance has positively shaped my research abilities immeasurably. I would like to extend thanks to other collaborators from whom I've developed a deeper appreciation for and understanding of developmental psychology and neuroscience: Dr. Joe Piven, Dr. Heather Hazlett, Dr. José Rodriguez-Romaguera, Dr. Casey Burrows, Dr. Kelly Caravella, Dr. Mark Shen, and Dr. Dea Garic.

On a more social level, I want to thank my friends and family, without whom these past several years would have been overly draining. My sincerest thanks to Nick Adams, Jared Maxwell, Jeremy and Emily Adler, Katie Stoffel, and Kevin Donovan for checking in with me, even when I was too focused to check back in a timely manner. In addition to my dad and younger brother, I also want to thank my aunt Peggy and uncle Tom for your emotional support.

Research reported in this dissertation was funded in part by NICHD of the National Institutes of Health under award number R01HD055741.

Last, but certainly not least, I want to thank my two ever-present research assistants, Edgewood and Beechwood.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

AIDS	Acquired Immune Deficiency Syndrome
ASD	Autism Spectrum Disorder
CART	Classification and Regression Tree
DTI	Diffusion Tensor Imaging
EL	Elevated Likelihood
EL-ASD	Elevated Likelihood with Autism Spectrum Disorder diagnosis
EL-Neg	Elevated Likelihood with no Autism Spectrum Disorder diagnosis
HARE	Hazard Regression
HIV	Human Immunodeficiency Virus
HL	High Likelihood
HL-ASD	High Likelihood with Autism Spectrum Disorder diagnosis
HL-Neg	High Likelihood with no Autism Spectrum Disorder diagnosis
JLCMM	Joint Latent Class Mixed Model
LL	Low Likelihood
LL-ASD	Low Likelihood with Autism Spectrum Disorder diagnosis
LL-Neg	Low Likelihood with no Autism Spectrum Disorder diagnosis
LMM	Linear Mixed-Effects Model
MARS	Multivariate Adaptive Regression Splines
MMSE	Mini-Mental State Examination
MRI	Magnetic Resonance Imaging
PH	Proportional Hazards
T4+	CD4 T-lymphocyte count
ZDV	Zidovudine
λ	Hazard
θ	Fixed-effects for Survival Model
β	Fixed-effects for Linear Mixed-Effects Model

b	Random-effects for Linear Mixed-Effects Model
ϵ	Random error
ω	Stationary Process
μ	Population Mean
σ	Population Standard Deviation
$\bar{x}$	Sample Mean
η	Latent Variable
γ	Shared Random Effect

CHAPTER 1: JOINT MODELS IN STATISTICS

1.1 The Origins of Joint Modeling in Biomedical Research

During the height of the AIDS epidemic in the United States in the late 1980's to mid 1990's, several serologic markers were found to be associated with the development of AIDS among those infected with HIV (Fahey et al., 1990; Schecter et al. 1989). One serological biomarker determined to have an association is the count of CD4 T-lymphocytes, more commonly known as T-cells. This discovery led to the development of zidovudine (ZDV) to treat HIV by inhibiting the replication of the HIV type 1 virus in vitro (Fischl et al., 1990). While ZDV worked through the suppression of HIV replication, the efficacy of this drug was tested by estimating CD4 T-lymphocyte (T4+) counts in blood samples. The primary issue with these early studies was that they were cross-sectional, estimating associations between T4+ counts and time to AIDS at a single point in time. This issue raised several difficulties. This methodology cannot tell whether a decline of T4+ count within an individual is associated with time to AIDS development. If HIV infection time was unobserved (i.e., left censoring was present) or time to AIDS after enrollment was unobserved (i.e., right censoring was present), then a cross-sectional method produces biased estimates. Even in a longitudinal or survival analysis setting, the estimation between T4+ count and time to AIDS was challenging to model because T4+ counts varied considerably between individuals and both types of censoring could be informative (Pawitan & Self, 1993).

One of the earliest attempts to model this association while accounting for these issues was done by De Gruttola, Lange, and Dafni (1991). They used longitudinal data and modeled the cohort's T4+ count using a growth model with random effects where measurement error was added as a weight to the random effects estimates. The added measurement error accounted

for the left censoring issue of not observing HIV infection time, whereas not accounting for this error could bias the estimated parameters towards the null (Prentice, 1982). As Pawitan and Self (1993) note, this method accounts for left censoring but not for the right censoring of unobserved development of AIDS. As such, Pawitan and Self developed a method to estimate jointly the time to HIV infection, time to AIDS development, and T4+ count where the former two components of the model were estimated with parametric survival models and the latter with a growth curve similar to De Gruttola, et al (1991). De Gruttola and Tu (1994) later improved the growth curve model from the original paper where observed times to AIDS development and T4+ count were jointly estimated assuming a multivariate normal distribution.

Wulfsohn and Tsiatis (1997) developed a joint method similar to the two aforementioned methods with less restrictive assumptions for both the survival and random effects models. The goals for the development of their method, however, stemmed from correcting drawbacks from a two-stage approach to modeling T4+ counts and time to AIDS development. The two-stage method estimated a growth curve with random effects for T4+ count and used the estimates from that model as covariates for a Cox model estimating the time to AIDS (Tsiatis, De Gruttola, & Wulfsohn, 1995). This approach makes strict assumptions that the random effects are normally distributed at each event time, which is unlikely. Drug abuse, for example, is highly associated with AIDS development through needle sharing (Des Jarlais & Friedman, 1988). A covariate representing drug use would have a large negative slope in its prediction of time to AIDS development leading to early development. The removal of these individuals from later random effects estimates would likely skew the distribution to a nonnormal one. A second concern stems from the lack of survival data in the growth curve model. While the survival model accounts for the covariates estimated by the growth curve, the converse is not true for this two-step approach. This setup treats survival data as secondary while making its usage less efficient in the model estimation. Faucett and Thomas (1996) proposed a Markov chain Monte Carlo technique with Gibbs sampling as a Bayesian alternative to jointly measuring T4+ count and time to AIDS development. This method was intended to improve on the two-stage method,

but it had the added benefit of an ability to treat covariate tracking and disease risk modeling with simultaneous interest. This latter benefit of the MCM model contrasted with the major limitation of the Pawitan and Self (1993) method, which conditioned time-to-AIDS on T4+ count thereby inhibiting the ability to make inference about T4+ count based on the time to development of AIDS.

The origin of joint models highlight two details of interest: the models were developed for subject-specific purposes and for accounting a variety of deficits endemic to the standard models. These joint models advanced the understanding of HIV progression into AIDS by improving the modeling of disease progression and of natural serologic history such as T4+ (Foulkes, 1998). Further, they were used to explore the efficacy of varying doses of ZDV against AIDS development and of aerosolized pentamidine against *Pneumocystis carinii* pneumonia common among those infected with HIV (Foulkes, 1998). Therefore, their utility for analyzing disease prognosis in tandem with biomarker trajectory emerges *prima facie*. Additionally, the focus on appropriately modeling the data, rather than attempting to transform data to fit within the standard statistical models, helped establish the theoretical framework of joint models. This foundation is apparent based on the history of joint models as improvements of two-stage and measurement error models. The subsequent question is how can these joint models be adapted for other types of data?

1.2 The Modern Joint Model for Longitudinal and Time-to-Event Data

While joint models encompass a wide array of models where distinct likelihood functions are simultaneously maximized, this manuscript focuses on joint models that include longitudinal and time-to-event data. These models have been applied to settings where participants are followed over time while tracking the repeated measurements of a biomarker to a disease as well as the time to an event related to that disease (e.g., onset of disease, death from disease, intervention, etc.) (Papageorgiou et al., 2019). For the sake of simplicity, the biomarker and time-to-event are two univariate dependent variables. However, extensions to multivariate

longitudinal outcomes (Lin et al., 2002a; Brown et al., 2005; Chi & Ibrahim, 2006; Rizopoulos & Ghosh, 2011) and to recurrent or competing events (Elashoff et al., 2008; Williamson et al., 2008; Hu et al., 2009) have been developed for joint models. Joint models offer several advantages over standard methods of analyzing a time-to-event outcome with a time-varying covariate (i.e., using the dynamic biomarker as an independent variable in the survival model). In a Cox model with a time-varying covariate, the value of this covariate is assumed constant in between measurements. This assumption leads the Cox model to produce biased estimates and standard errors (Prentice, 1982) as well as inefficient estimates (Ibrahim, Chu, & Chen, 2010). These issues derive from the endogeneity of the dynamic biomarker, where the value changes over time based on a set of other independent variables that vary randomly between participants. Joint models provide another advantage over simple survival models because of their ability to make inference on the joint distribution of the longitudinal outcome and the censoring/missingness of the time-to-event outcome (Rizopoulos, 2012). The probability of a participant's dropout from a study may depend on the level of the biomarker prior to dropout. Therefore, the censoring of the event is informative and cannot be ignored. Joint models are required when the study's interest is understanding how the repeated biomarker measurement and time-to-event are linked (Proust-Lima et al., 2014). Finally, joint models provide a superior framework for prediction compared to the standard model due to the presence of dynamic components within the joint model (Proust-Lima et al., 2014; Król et al., 2017).

The joint model comprises three primary components: (i) a model for the time-to-event outcome, (ii) a model for the biomarker trajectory, and (iii) a method for linking the two component structures. Component (i) is typically modeled as a survival model, with the literature holding particular interest in either parametric models such as Weibull or exponential survival models (Ibrahim, Chu, & Chen, 2010) or in a proportional hazards model such as the Cox proportional hazards model (Proust-Lima et al., 2014). Component (ii) primarily uses a generalized linear mixed-effects model–more specifically a linear mixed-effects model–to model the biomarker trajectory. The method for component (iii) has produced a fork in the

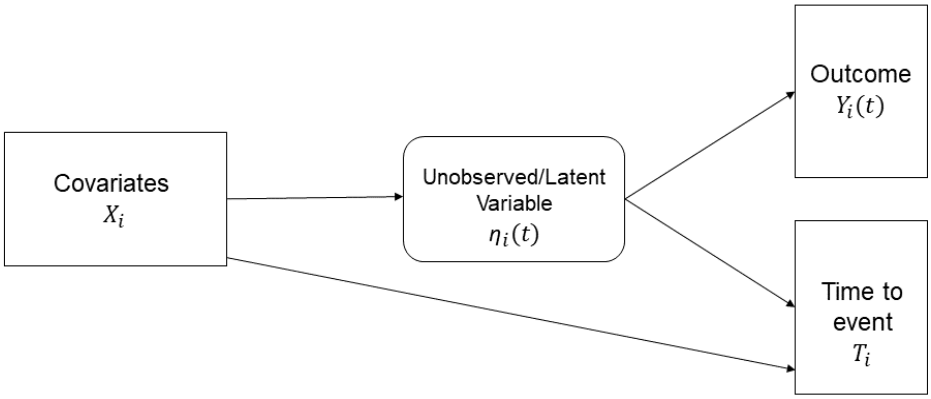

Figure 1.1.1 Diagram of a Joint Model

current literature producing two general ways of approaching this linking mechanism. Prior to delineating these details, however, first consider the model shown in Figure 1.1.1 that illustrates the causal path underlying the joint model.

Much of the literature describes the longitudinal outcome as a biomarker (Proust-Lima et al., 2014; Papageorgiou et al., 2019). This description, while useful shorthand, may be imprecise as various types of longitudinal outcomes may be modeled in a joint model such as quality-of-life scores or health outcomes (Ibrahim, Chu, & Chen, 2010) or psychometric scores (Xu and Zeger, 2001; Proust-Lima et al., 2017). Pawitan and Self (1993), Wulfsohn and Tsiatis (1997), and Faucett and Thomas (1996) all describe joint models where the observed serologic biomarker represents an unobserved health status. That is, T4+ count reflects the infection prognosis of HIV. Xu and Zeger (2001) make this point more explicit by noting how joint models of this type are latent variable models that estimate T_i, Y_i given independent variables X_i as $[T, Y|X] = \int [T, Y|\eta, X] d[\eta|x]$. The estimates of the time-to-event and the

longitudinal outcome are determined through the change in some unobserved variable, η. The goal of the studies described in §1.1 was not to estimate how T4+ count affects time to AIDS development, but rather to estimate how HIV infection status affects time to AIDS development where this status is indirectly observed through T4+ count. Any number of serologic or immunologic biomarkers could have been used as the longitudinal outcome as long as that biomarker appropriately represented the latent variable of HIV infection status. The manner in which component (iii) is determined in a joint model largely dictates which of the prongs from the aforementioned fork a researcher takes in modeling T and Y.

Xu and Zeger (2001) elaborate on the formulation of the joint estimation of T, Y as $[T, Y|X] = \int [T, Y|\eta, X][\eta, X]d\eta = \int [T|\eta, X][Y|\eta][\eta|X]d\eta$. This equation essentially restates the information provided in Figure 1.1.1. From the equation and diagram, three assumptions for this latent variable method of joint modeling appear:

(a) Given knowledge about η, the latent variable, T is independent from Y,

(b) X affects T either directly or indirectly through η,

(c) X only affects Y through the latent variable η.

These assumptions highlight the importance of how η is treated in the joint model. This latent variable either directly wholly or partially determines the two outcomes, T and Y. Two popular methods for the treatment of η is the Shared Random Effects Model and the Joint Latent Class Mixed Model.

1.3 The Shared Random Effects Model

The Shared Random Effects Model (SRM) treats η through the use of a common random effect between the longitudinal model and time-to-event model where the random effect from the former is incorporated as a covariate in the latter. The longitudinal outcome, $Y_i(t)$, is assumed to be the true latent variable $\eta_i(t)$ plus some error $\epsilon_i(t)$ or $Y_i(t) = \eta_i(t) + \epsilon_i(t)$. In practice, this equation is often modeled using a linear mixed-effects model with $\eta_i(t) = b_{0i} + b_{1i}(t) + \beta X_i$

where b_{0i}, b_{1i} are multivariate normal random effects and β is the fixed effect (Ibrahim, Chu, & Chen, 2010). Note that additional fixed and random effects may be added to this model. The time-to-event model in practice is often modeled with a parametric survival model such as the Weibull distribution, and the hazard function is estimated $h(t) = h_0(t)\exp(\gamma\eta_i(t) + \alpha X_i)$ where $h_0(t)$ is the baseline hazard function, γ is the effect of the longitudinal model on the time-to-event, and α is the effect of the covariates on the time-to-event (Ibrahim, Chu, & Chen, 2010). Using the language of structural equation modeling (Bollen, 1989), α is the direct effect of the covariates on the time-to-event, β is the indirect effect of the covariates on the time-to-event, and $\gamma\beta + \alpha$ is the total effect. If $\gamma = 0$, then there is no association between the longitudinal and survival models, and a joint model does not improve the estimation of the time-to-event compared to a survival model alone (Ibrahim, Chu, & Chen, 2010). The function of the random effects from the longitudinal model can take many forms in the survival model such as individual deviation from the expected longitudinal trajectory, the instantaneous rate of the trajectory at given time points (i.e., the derivative of the trajectory), or some combination of functions (Proust-Lima et al., 2014).

The SRM is an appropriately popular joint model due to the simplicity of its inference. The parameters are linear and separable in their estimation, and the effects of each outcome on one another can be determined through a composition of their respective parameters. The estimates of the parameters are more efficient and less biased than survival models with time-varying covariates (Ibrahim, Chu, & Chen, 2010). Further, joint models typically lead to smaller standard errors. These properties have implications for study design, where a study using a joint model can require smaller sample sizes for equivalent power to a study using only a longitudinal or survival model. Proust-Lima et al. (2014) note that SRMs are suitable joint models when the researcher wants to test specific assumptions regarding characteristics of the longitudinal trajectory on the time-to-event. That is, SRMs work well under circumstances when the relationship between Y and T are either well-understood or the hypothesis test for one outcome conditioned on another is specifically defined.

The SRM has received extensive attention in the literature due to these aforementioned properties. Ibrahim, Chu, & Chen (2010) detail a SRM of a cancer treatment trial that jointly estimated the Functional Assessment of Cancer Therapy–Breast Cancer (FACT-B) quality of life scale (Brady et al., 1997) and overall survival time. The treatment compared doxorubicin, paclitaxel, and a combination of the two jointly on the FACT-B quality of life score, where a higher score indicated greater quality of life, and the total survival time after the beginning of treatment administration. The SRM estimated the overall treatment hazard ratio of 0.76 whereas the hazard ratios for a Cox model without the trajectory estimate and a two-stage model were 0.78 and 0.77, respectively. These results corroborated with Prentice (1982) that two-stage models attenuate treatment effect sizes. Further, these results indicated that a joint model provided a less biased estimate of the treatment effect compared with the simple Cox model. Long & Mills (2018) used a SRM to jointly model a multivariate longitudinal outcome comprising cytosine-adenine-guanine (CAG) trinucleotide mutation in the *HTT* gene of chromosome 2, a total motor score (TMS), and a symbol digit modalities test score (SDMT) along with the time-to-Huntington's Disease onset. These models were used to estimate individual-specific predictions of time-to-Huntington's Disease probabilities. This model was compared across four different studies. The authors found that the SRM had better predictive ability and external validity than traditional Cox models for estimating the time-to-event. Similar results were found for TMS score prediction across the four studies and mixed results were found among the SDMT results.

The implicit assumption of η within a SRM is that it is homogeneous within the sample. That is, the latent variable follows a single distribution of which all participants deviate from a single mean. This assumption highlights a limitation with the SRM. As noted by Ibrahim, Chu, & Chen (2010), a common trajectory structure is assumed for treatment groups. For example, if one group has an effect that is linear in time whereas another group has an effect that is quadratic in time, the SRM would not suitably model the data. More succinctly, heterogeneity in η poses problems for the traditional SRM.

1.4 The Joint Latent Class Mixed Model

While receiving less attention than the SRM in the current literature (Proust-Lima et al., 2014), the Joint Latent Class Mixed Model (JLCMM) provides an alternative method of joint modeling that is able to handle some of the limitations imposed on the SRM. The popularity of the JLCMM derives from its ability to handle heterogeneity in η, which is accomplished by estimating k latent classes that compose a mixture distribution for η. The methods for estimating the longitudinal and time-to-event outcomes are the same as those used in a SRM, excepting the estimation of a parameter that relates these two outcomes to one another, and the determination of latent classes is estimated with multinomial logistic regression. Further, the longitudinal and survival models in a JLCMM may include class-specific estimates. Details about the JLCMM are provided in chapter 3.

The JLCMM distinguishes itself from the SRM in several ways. The primary distinction is how the relationship between the longitudinal outcome Y and time-to-event outcome T is modeled. The SRM uses random effects to model within-subjects covariance as well as the dependence between Y and T whereas the JLCMM uses random effects only to model within-subjects covariance. The JLCMM accounts for dependence between Y and T by assuming that determination of the joint outcome $[Y, T]$ depends on the membership of a latent class. That is, once latent class k is known, $[Y, T]$ can be estimated. This distinction between the SRM and JLCMM illustrates another distinction: the SRM requires variance in the population to be homogeneous (i.e., from a single distribution) whereas the JLCMM assumes that the variance is heterogeneous (i.e., from a mixture of several distributions). The homogeneity of the SRM, while a simpler assumption, is also stricter. The trajectory of the longitudinal outcome and its functional form in the time-to-event model are typically chosen *a priori* (Proust-Lima et al., 2014), and such a decision affects the parameter estimation and model fit. Further, the functional form assumptions can be too strict when the goal is predictive ability. The JLCMM instead has less strict assumptions with its latent class characterization of η by not assuming any functional dependence between Y and T. This characterization provides more flexibility by allowing

class-specific estimates of parameters from both the longitudinal and survival models. However, this method also has a cost of increasing the total number of parameters to be estimated in the joint model (Proust-Lima et al., 2014). One final, technical difference between the two methods relates to the numerical estimation of the log-likelihood function. The SRM requires numerical integration over the distribution of random effects. The JLCMM replaces numerical integration with summing over the latent classes. This latter operation is less computationally intensive than the former (Proust-Lima et al., 2014).

One of the earliest uses of the JLCMM was estimating the utility of prostate-specific antigen (PSA) as a biomarker for onset of prostate cancer (Lin et al., 2002b). Lin et al. (2002b) were motivated by early detection of prostate cancer as its survival probability increases the earlier prostate cancer is detected. The evaluation of blood PSA levels had been a promising diagnostic tool for prostate cancer due to prostate cancer having few early stage symptoms (Catalona, Smith, & Ornstein, 1997). The decision to use a JLCMM rather than a SRM for this hypothesis stems from the heterogeneous expression of normal blood PSA and its divergent outcomes. While 3 - 4 nanograms per milliliter of blood PSA is normal, approximately 20% of men diagnosed with prostate cancer had PSA levels within that range (Catalona, Smith, & Ornstein, 1997). Further, high blood PSA levels can indicate other more likely diseases such as benign prostatic hyperplasia or prostatitis (Lin et al., 2002b). Prostate cancer also has different disease patterns where its onset either follows a slow progression or a much more aggressive one (Morell et al., 1995), perhaps due to genetic factors (Padilla-Nash et al., 2001). Unlike the T4+ counts from the HIV-to-AIDS analyses from §1.1, blood PSA levels have heterogeneous η where the joint outcomes $[Y, T]$ follow different patterns based on unobserved factors. By using the JLCMM, Lin et al. (2002b) estimated four latent classes corresponding to those with very low levels of PSA, low levels of PSA, medium levels of PSA, and high levels of PSA in their blood. Within these classes, age was positively associated with prostate cancer onset and selenium levels were negatively associated with prostate cancer onset. As a comparison, Lin et al. calculated four separate longitudinal and survival trajectories based on baseline blood

PSA levels that paralleled the four latent classes. The latent classes tracked these four pairs of observed curves very well, but the JLCMM discriminated among these curves better than the empirical model (Lin et al., 2002b). That is, the JLCMM determined the four classes without prior assumptions and better modeled the data than a model that would make such assumptions.

Garre et al. (2008) jointly modeled reciprocal serum creatinine concentration (RC) and time-to-kidney graft failure using a JLCMM. The primary scientific distinction between this model and the one developed by Lin et al. (2002b) was that the number of latent classes was known *a priori*. RC patterns typically follow two patterns: a stable pattern typically precedes no observed graft failure and a sudden drop in RC values typically precedes observed graft failure. However, this timing in RC drop is not easily predictable, hence the utility in using a JLCMM to help predict RC trajectory and renal graft failure time (Garre et al., 2008). Given a set of risk factors as class-specific covariates, the JLCMM predicted failure within each 2-year interval of the study much better than a SRM that assumed linear random effects (Garre et al., 2008). In certain cases where some prior knowledge of the population is known, a JLCMM may still be preferable for joint modeling than a SRM.

Proust-Lima et al. (2014) describe two advantages for a JLCMM over a SRM. First, a JLCMM should be used when developing a predictive joint model. Additionally, a JLCMM should be used when investigating the link between a longitudinal marker and time-to-event without specific assumptions, particularly in cases where η is heterogeneous. These scenarios have largely been confirmed in the biomedical literature, particularly with disease prognosis. However, heterogeneity occurs often in other scientific fields where such joint models may prove useful. Psychology and neuroscience are both rich in longitudinal and time series data where the participants being tracked over time likely follow heterogeneous patterns of η. Therefore, JLCMMs have substantial promise in the modeling of these data and testing of complex hypotheses.

1.5 Heterogeneity in Psychology

The heterogeneity of variance in psychology, and by extension in neuroscience and neuroimaging, has long been recognized as inherent to the data and in need of careful statistical attention (Bryk & Raudenbush, 1988; Grissom, 2000; Lubke & Muthén, 2005; Masyn, 2013). No single source causes this heterogeneity, and it can derive from design factors within the study protocol (Hanselman et al., 2017), unidentified moderators or third variables (Aguinis & Pierce, 1998), or a genuine characteristic of the study population, for example. One common source of heterogeneity in the behavioral assessment of those with neurogenetic disorders is comorbidity. For example, those with Williams syndrome have increased risk for anxiety (Dykens, 2003), and people with Down syndrome have an elevated likelihood of being diagnosed with autism spectrum disorder (Oxelgren et al., 2017). Many researchers of neurodevelopmental disorders highlight the necessities and advantages of examining the heterogeneity within their study populations (Grzadzinski, Huerta, & Lord, 2013) with particular focus on latent profile analysis (Fider et al., 2022; Prince & Fidler 2021), latent class analysis (Adjacic-Gross et al., 2018), and finite mixture modeling (Masyn, 2013). Fidler et al. (2022) highlight the need to use "person-centered and data-driven approach[es]" in order to identify latent profiles of children with Down syndrome without the use of *a priori* knowledge.

The data-driven motivation applies largely to mixture modeling and latent class determination. By design, latent classes or profiles are constructed with few if any assumptions (Masyn, 2013). This construction is apparent in a JLCMM when estimating the latent class membership of each participant. The survival model of a JLCMM can be semiparametric, which aligns with having few assumptions about the data's distribution. However, even semiparametric models–such as the Cox proportional hazards model–have some assumptions that conflict with a latent class model. For example, the proportional hazards assumption (more details in Chapter 2) requires the hazard function to be proportionate among groups, which is unlikely given the nature of heterogeneity. A data-driven/adaptive method would likely better suit the survival model of a JLCMM.

1.6 An Aside on Adaptive Methods

From the 1950s to the 1990s, the U.S. federal government provided relatively vast amounts of funding for survey research in order to monitor the effects of infrastructural and systemic investsments that it made within the U.S. population (Groves, 2011). Morgan and Sonquist (1963) recognized several analytical issues that arose from the sudden availability of "rich data from cross section survey[s]". These issues encompassed various concepts well-known in the modern literature, including concerns about the vast amount of information for each participant (an *Ur* form of big data, perhaps), the representation of continuous variables as categorical, the inherent measurement error in all variables, the sophisticated sampling structures, the collinearity among many of the surveyed variables, and the temporal absence of cross-sectional data. Without providing a comprehensive delineation of how many of these concerns have been addressed since the inception of this paper, the measurement and collinearity issues are typically handled with latent variable methods such as structural equation modeling (Bollen, 1989) and temporality has been addressed with longitudinal data analysis (Diggle, Liang, & Zeger, 1994). Morgan and Sonquist (1963) in their paper focused on one particular issue which either predicates or relates to the others: a problem of "interaction effects". Specifically, their concern was that in the social sciences two effects are typically not additive, but rather have a stronger effect in tandem than as independent components. Statisticians often use interaction terms in their regression models to account for this. Many social scientists have particular interest in these effects with respect to concepts of mediation and moderation (Baron & Kenny, 1986). Further, Morgan and Sonquist note that these interactions may be inherent to the theoretical construct itself. A family life cycle, for example, can be considered an interaction between age, marital status, presence, and the age of children. That is, the family cycle is a latent variable expressed by these observations provided in survey data.

Morgan and Sonquist do not use the term "latent variable" to describe this construct, but it largely fits their description of the theoretical construct. To overcome many of the shortcomings of ordinary least squares regression for analyzing these types of data, they established a method

to identify subgroups that compose a latent variable that existed in the data. As they noted, "Perhaps the most important thing to keep in mind about survey data in the social sciences is that the theoretical constructs in most theory are not identical with the factors we can measure in the survey" (Morgan & Sonquist, 1963). This method was an adaptive one with 'adaptive' meaning adapting with the data or data-driven. While using the whole of the data of interest, their analysis would split the data using a dichotomous variable into two groups that explained the largest sum of squares. From the two groups, two more groups would be produced similarly. This method continued until any further subdivision failed to explain an additional 1% or more of the residual sum of squares. The branching paths from these variables would compose the subgroups of interest, aiding in understanding the relationship between a dependent variable and several interactions of independent variables. Morgan and Sonquist's method was generalized into Classification and Regression Trees (CART) by Breiman, Friedman, Olshen, and Stone (1984). CART uses an adaptive method to make decision rules that split independent variables X into intervals that best predict a dependent variable Y. CART uses different objective functions from Morgan and Sonquist's method, such as entropy or the Gini index, to evaluate the fit of the decision rule. However, the goal is ultimately the same: produce a set of homogeneous subgroups to explain the heterogeneity in the dependent variable.

CART influenced several other adaptive methods including additive modeling (Friedman & Silverman, 1989), multivariate adaptive regression splines (Friedman, 1991), and hazard regression (Kooperberg, Stone, & Truong, 1995a). Chapters 2 & 3 will contain more details about hazard regression. To summarize, both multivariate adaptive regression splines and hazard regression are regression methods that make minimal assumptions between the dependent and independent variables. In a stepwise fashion, the model adds knots along a range of values for the independent variables and adaptively chooses knot location based on various objectives such as minimizing the residual sum of squares. The final model is chosen based on maximum fit to the data. The important detail from these methods is that they are estimated with respect

to how the data are distributed rather than by assuming the data follow a specific probability distribution.

The JLCMM estimates several homogeneous subgroups from a heterogeneous sample. As such, it can be difficult to make assumptions about what latent classes exist and how their $[Y, T]$ are distributed. Therefore, any method that can construct a component model of the JLCMM solely from data-driven methods has advantages to methods that require functional or parametric assumptions. Latent classes will typically have functionally different joint trajectories of $[Y, T]$, and these different trajectories motivate the estimation of class-specific parameters using data-driven methods. The goal of this manuscript is to illustrate the use of such an adaptive method, hazard regression, to estimate the hazard function of a JLCMM. This hazard function will be used to estimate T in the JLCMM. Further, the class-specific hazard functions will be estimated adaptively helping to account for heterogeneity within the data. This method will have utility in psychology and neuroscience research, which is rich with such heterogeneity.

CHAPTER 2: INTRODUCTION TO METHODS AND CONCEPTS

2.1 Overview

The following chapter provides introductions to several methods and concepts that compose the two primary topics of this manuscript: joint latent class mixed models (JLCMMs) and hazard regression (HARE). The following methods and concepts are not described exhaustively. Rather, each section provides a sufficient summary to allow for understanding how each topic relates to JLCMMs, HARE, or the analyses completed to justify the method proposed in this manuscript. This chapter is structured in the following order: summaries of statistical models (§2.2 - §2.4) are followed by summaries of adaptive methods (§2.5 - §2.8) which are followed by summaries of the substantive topics discussed in the manuscript (§2.9 - §2.10). Each section is mostly self-contained with clear references made to other sections when necessary. This chapter is a reference for other chapters. Therefore, the reader may choose only to read sections of interest or of necessity. Sections not of interest may be skipped without hindering the comprehension of the rest of the manuscript.

2.2 Latent Class Analysis

Conditional Probability and Bayes' Theorem

It is often useful to calculate the probability of event X while assuming some other event Y. This probability, $P(X|Y)$, is denoted as the conditional probability of X given Y. Formally this probability is defined as

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} \tag{2.1}$$

where $\cap$ is the intersection symbol, effectively meaning "and". Finding the probability of X while assuming Y is described as finding the probability of X and Y concurrently while considering only cases where Y is true. Unless X and Y are independent, then knowing something about event Y improves the determination of the probability of event X. Bayes' Theorem uses conditional probability centrally, often considering the calculated probability as a "posterior probability" based on updating a prior belief with new evidence. Formally Bayes' Theorem is defined as

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)}. \tag{2.2}$$

Bayes' Theorem is used often in evaluating diagnostic tests. For example, if X is the event of a positive test result and Y is the event of having a disease, then the probability of having a disease given a positive test result is then seen as the probability of having a positive test given having said disease ($P(X|Y)$) times the probability of having said disease at all ($P(Y)$) among all cases of having a positive test result ($P(X)$). The conditional probability of $P(Y|X)$ is an improvement over the simple probability of having the disease $P(Y)$ since the positive test gives more information about disease status.

Estimating Latent Classes

Given a statistical model, assume that some parameters within the model vary between unobserved groups. The identification of these unobserved groups based on the values of these parameters forms the basis of latent class analysis. Let C represent one of these unobserved groups–a latent class–and let X represent an observed variable in the data. Further, let there be $k = 1, \ldots, K$ latent classes and all participants in the study $i = 1, \ldots, n$ belong to a latent class. For example, $c_i = 3$ implies that person i belongs to class 3. The goal of latent class analysis (LCA) is to determine c_i for all i given the responses to each variable in X. To illustrate this point, consider Table 2.1.1 which displays simulated data. X_1, X_2, X_3 are all dichotomous variables taking values of 0 or 1. These variables can be seen as survey responses of "No" or "Yes" corresponding to 0 and 1, respectively.

Table 2.2.1 Fictional LCA Data

X_1	X_2	X_3	N	$P(C=1\|\mathbf{X}=\mathbf{x})$	$P(C=2\|\mathbf{X}=\mathbf{x})$	$P(C=3\|\mathbf{X}=\mathbf{x})$
0	0	0	500	0.995	0.003	0.002
0	0	1	77	0.821	0.178	0.001
0	1	0	212	0.876	0.113	0.011
0	1	1	16	0.432	0.157	0.411
1	0	0	107	0.710	0.187	0.103
1	0	1	439	0.357	0.120	0.523
1	1	0	62	0.106	0.811	0.083
1	1	1	403	0.001	0.001	0.998

Different response patterns correspond to different latent classifications. For example, a response of all 0s implies a 99.5% probability of being a member of latent class 1. Endorsing 0 only for X_2 implies a probability of 12% of being a member of latent class 2, and so forth. The law of total probability indicates that over m partitions,

$$\begin{aligned} P(X) &= \sum_m P(X \cap Y_m) \\ &= \sum_m P(X|Y_m)P(Y_m) \end{aligned}$$

where $\cap$ is the intersection between event spaces. In this example, the probability of a given response pattern for X is a weighted sum of the probability of being in latent class C times the conditional probability of responding to the X values in a specific way given being a member of latent class C. That is,

$$P(\mathbf{X}=\mathbf{x}) = \sum_{k=1}^{K} P(C=k)P(\mathbf{X}=\mathbf{x}|C=k) \tag{2.3}$$

In equation (2.3), the probability $P(C=k)$ is simply the proportion of individuals who belong to latent class k. One of LCA's assumptions is that responses to X are independent within each

latent class. In other words,

$$P(\mathbf{X} = \mathbf{x}|C = k) = \prod_{l=1}^{L} P(X_l = x_l|C = k). \tag{2.4}$$

Combining equations (2.3) and (2.4) implies

$$P(\mathbf{X} = \mathbf{x}) = \sum_{k=1}^{K} P(C = k) \prod_{l=1}^{L} P(X_l = x_l|C = k). \tag{2.5}$$

Table 2.2.2 provides probabilities for responses to each item conditioned on the latent class. These values have been calculated directly from Table 2.2.1. For example, to find the probability (or proportion) of being a member of latent class 1:

$P(\text{Member of latent class 1}) =$

$P(\text{Member of latent class 1 \& responded } X_1 = 0, X_2 = 0, X_3 = 0)$ OR

$P(\text{Member of latent class 1 \& responded } X_1 = 0, X_2 = 0, X_3 = 1)$ OR

$\ldots$ OR $P(\text{Member of latent class 1 \& responded } X_1 = 1, X_2 = 1, X_3 = 1)$

$$\begin{aligned}
= P(C = 1) &= \sum^{X_1, X_2, X_3} P(C = 1|\mathbf{X} = \mathbf{x}) \times P(\mathbf{X} = \mathbf{x}) \\
&= \frac{1}{1816}(500 \times 0.995 + 77 \times 0.821 + 212 \times 0.876 \\
&\quad + 16 \times 0.432 + 107 \times 0.710 \\
&\quad + 439 \times 0.357 + 62 \times 0.106 + 403 \times 0.001) \\
&\quad = 0.55.
\end{aligned}$$

Further the probability of responding 0 to X_2 given someone is a member of latent class 3 ($C = 3$) is:

$$\begin{aligned} P(X_2 = 0|C = 3) &= \frac{P(X_2 = 0 \cap C = 3)}{P(C = 3)} \\ &= \frac{1}{0.36} \frac{(500 \times 0.002 + 77 \times 0.001 + 107 \times 0.103 + 439 \times 0.523)}{1816} \\ &= 0.3697 \approx 0.37 \end{aligned}$$

Table 2.2.2 Response Probabilties of Fictional LCA Data

	$C = 1$	$C = 2$	$C = 3$
$P(C = k)$	0.55	0.09	0.36
$P(X_1 = 0\|C = k)$	0.76	0.25	0.02
$P(X_2 = 0\|C = k)$	0.80	0.53	0.37
$P(X_3 = 0\|C = k)$	0.77	0.58	0.03

If the probabilities of response profiles conditioned on latent classes is known (i.e. $P(\mathbf{X} = \mathbf{x}|C = k)$), then the probability of latent class membership can be determined using Bayes' Theorem (2.2).

A Loglinear Formulation of LCA

Haberman (1979) and Vermunt & Magidson (2004) illustrated that a latent class model could be specified as a log-linear model. That is,

$$\ln P(C = k, \mathbf{X} = \mathbf{x}) = \xi + \xi_k^C + \sum_{l=1}^{L} \xi_{x_l}^{X_l} + \sum_{l=1}^{L} \xi_{k,x_l}^{C,X_l} \tag{2.6}$$

where the ξ are the ANOVA terms of the model for the main effect, latent class effect, response profile effect, and interaction between the two. The effects including the observed response profiles are summed over the levels of the responses and their sums total to 0 (i.e. $\sum_{l=1}^{L} \xi_{x_l}^{X_l} = \sum_{l=1}^{L} \xi_{k,x_l}^{C,X_l} = 0$) for purposes of identifiability.

This formulation becomes important for a few reasons. Primarily, the log-odds of one response profile compared to another is additive using this formulation. For a two-by-two contingency table, where μ_{ij} is the count in cell with row i and column j, the log-odds is calculated by

$$\begin{aligned}
\ln(\theta) &= \ln \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}} \\
&= \ln \mu_{11} + \ln \mu_{22} - \ln \mu_{12} - \ln \mu_{21} \\
&= (\xi + \xi_1^A + \xi_1^B + \xi_{11}^{AB}) + (\xi + \xi_2^A + \xi_2^B + \xi_{22}^{AB}) \\
&- (\xi + \xi_1^A + \xi_2^B + \xi_{12}^{AB}) - (\xi + \xi_2^A + \xi_1^B + \xi_{21}^{AB}) \\
&= (\xi + \xi - \xi - \xi) + (\xi_1^A - \xi_1^A + \xi_2^A - \xi_2^A) \\
&+ (\xi_1^B - \xi_1^B + \xi_2^B - \xi_2^B) + (\xi_{11}^{AB} + \xi_{22}^{AB} - \xi_{12}^{AB} - \xi_{21}^{AB}) \\
&= 0 + 0 + 0 + \xi_{11}^{AB} + \xi_{22}^{AB} - \xi_{12}^{AB} - \xi_{21}^{AB} \\
&= \xi_{11}^{AB} + \xi_{22}^{AB} - \xi_{12}^{AB} - \xi_{21}^{AB}.
\end{aligned}$$

Therefore, a change in the probability of latent class membership can be simply estimated from a change in the response profile. Further, this formulation can be constructed into a multinomial logit model where latent class membership can be estimated from a set of Z factors, where Z can be either a response profile for a set of items, covariates of interest, or both. More specifically, the latent class membership probability is

$$P(C = k|\mathbf{Z}) = \frac{\exp\left(\xi_{0k} + \mathbf{Z}^T\xi_{1k}\right)}{\sum_{g=1}^{K} \exp\left(\xi_{0g} + \mathbf{Z}^T\xi_{1g}\right)}. \tag{2.7}$$

Each latent class k has its own estimates, an intercept ξ_{0k} and set of coefficients ξ_{1k} for the design matrix $\mathbf{Z}$. Equation (2.7) is ultimately used to estimate latent class membership in the JLCMM.

Estimation of Latent Classes

One implicit assumption made throughout §2.2 was that latent class membership was known *a priori* to data collection, thus allowing for estimating its association with response profiles. In practice, this is not the case. Rather, latent classes are undetermined or missing, and the goal of a LCA is to estimate how many latent classes there are from the data. Therefore, an analytic solution to this question is unfeasible, and a numerical solution is required. Many numerical methods can be used for LCA, and the primary algorithm used throughout this manuscript is the Levenberg-Marquardt algorithm (§2.8). However, the EM algorithm will be discussed here for simplicity.

Assume a normal mixture model composed of multiple normal distributions, $N_j \sim (\mu_j, \sigma_j^2)$ for $j = 1, \ldots, J$ total clusters of normal distributions. Suppose that $\mathbf{X} = (X_1, \ldots, X_n)$ are the data from the sample. Assume $\mathbf{X}$ are independently and identically distributed random variables with a probability density function following

$$f_x(x) = \sum_{j=1}^{J} p_j \frac{1}{\sqrt{2\pi\sigma_j}} e^{-(x-\mu_j)^2/2\sigma_j^2}$$

The familiar part is the normal probability density function with parameters μ_j, σ_j^2. The "new" part is the probability parameter p_j, where $p_j \geq 0$ for all j clusters, and $\sum_j^J p_j = 1$. The goal is to determine the total number of clusters/latent classes J by estimating p_j, μ_j, σ_j^2.

The EM algorithm works by assuming some data are missing from the sample. This method is useful for determining the number of latent classes because the latent class is treated as missing data, which makes sense considering it is unobserved. The method comprises two steps: (i) a step that find the expected value of the log-likelihood with respect to the missing data (E-step), (ii) a step that maximizes this expectation over the current parameter estimates (M-step). After each M-step, the most recent parameter estimates are put into the E-step to start the process again. This process works iteratively until the old estimate and new estimate are negligibly different.

For this problem, let C represent the variable corresponding to the latent class. Then $P(C = j) = p_j$ and a member m belonging to class j will be noted as coming from a normal distribution with mean μ_j and variance σ_j^2. Let $(C_1, \ldots, C_J) := \mathbf{C}$. Then the conditional probability density of the data X given membership in class $\mathbf{C}$ is

$$f_{x|\mathbf{C}}(x_i|c_i = j, p_j, \mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j}} e^{-(x-\mu_j)^2/2\sigma_j^2}$$

with given likelihood function

$$L(p_j, \mu_j, \sigma_j^2, \mathbf{X}, \mathbf{C}) = \prod_{i=1}^{n} p_{c_i} \frac{1}{\sqrt{2\pi\sigma_{c_i}}} e^{-(x-\mu_{c_i})^2/2\sigma_{c_i}^2}.$$

Let $\theta = (\mathbf{p}, \mu, \sigma^{\mathbf{2}})$ represent the set of parameters in the model that need to be estimated by the EM algorithm. Further, let $\theta^{(k)}$ be the current estimates and $\theta^{(k+1)}$ be the new estimates. In practice, the first estimate of $\theta^{(k)}$ is guessed, where a good guess can help with convergence. Then the EM algorithm for this problem proceeds as such:

E-Step: Define $Q(\theta^{(k+1)}; \theta^{(k)}) := E[\ell(\theta; \mathbf{C}, \mathbf{X})|\mathbf{X}, \theta^{(k)}]$. That is, Q is the function that finds the expected value of the log-likelihood function given the data and the current estimates of the parameters. Since Q is an expectation, its calculation is straightfoward.

$$Q(\theta^{(k+1)}; \theta^{(k)}) = \sum_{i=1}^{n} \sum_{j=1}^{J} P(C_i = j|x_i, \theta^{(k)}) \ell(f_{x|\mathbf{C}}(x_i|c_i = j, \theta^{(k)}) \times P(C_i = j|\theta^{(k)})).$$

Using Bayes' Theorem and algebra, $P(C_i = j|x_i, \theta^{(k)})$ can be calculated (the other parts of $Q(.)$ are found directly from the initial parameter guess):

$$\begin{aligned} P(C_i = j|x_i, \theta^{(k)}) &= \frac{P(C_i = j \cap X_i = x_i|\theta^{(k)})}{P(X_i = x_i|\theta^{(k)})} \\ &= \frac{f_{x|\mathbf{C}}(x_i|c_i = j, \theta^{(k)})P(C_i = j|\theta^{(k)})}{\sum_{l=1}^{J} f_{x|\mathbf{C}}(x_i|c_i = l, \theta^{(k)})P(C_i = l|\theta^{(k)})}, \end{aligned}$$

which can be computed similarly to how the fictional data from Table 2.2.1 was calculated.

M-Step: Maximize $Q(\theta^{(k+1)}, \theta^{(k)})$ by solving the partial derivatives with respect to each parameter and set them equal to 0.

$$\begin{aligned} \mu_j^{(k+1)} &= \frac{\sum_{i=1}^{n} x_i P(C_i = j|x_i, \theta^{(k)})}{\sum_{i=1}^{n} P(C_i = j|x_i, \theta^{(k)})} \\ \sigma_j^{(k+1)} &= \frac{\sum_{i=1}^{n} (x_i - \mu_j)^2 P(C_i = j|x_i, \theta^{(k)})}{\sum_{i=1}^{n} P(C_i = j|x_i, \theta^{(k)})} \\ p_j^{(k+1)} &= \frac{1}{n} \sum_{i=1}^{n} P(C_i = j|x_i, \theta^{(k)}) \end{aligned}$$

Set $\theta^{(k)} = \theta^{(k+1)}$ and go through the E-step and M-step iteratively. A stopping rule, such as the absolute difference $|\theta^{(k+1)} - \theta^{(k)}| < \epsilon$, indicates when the algorithm ends.

2.3 Linear Mixed Effects Models

The Model

It is common to extend the ordinary least squares (OLS) regression model to include the estimation of "random effects" b in conjunction with the typical "fixed effects" β. The mainstream distinctions of fixed and random effects are somewhat nebulous. Gelman (2005) details five different definitions of these terms from various statistical literature, and they are

only loosely related. Rather than try to provide a definition for each effect type, this section will justify the utility of estimating a linear model composed of fixed and random effects. From this justification some distinction between the two sets of estimates can be given.

Assume $\mathbf{Y}$ is a vector of random variables of interest for $i = 1, \ldots, N$ participants. Let y_i represent these variables for participant i, where each participant can have a different number of variables, n_i. For example, if Y represents IQ scores over four time periods, n_i can be 0, 1, 2, 3, or 4 observations. The linear mixed effects model (LMM) is estimated as:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}b + \epsilon \tag{2.8}$$

where $\mathbf{X}$ is a design matrix of variables for the fixed effects, $\mathbf{Z}$ is a design matrix of variables for the random effects, β are the coefficients of the fixed effects, b are the coefficients of the random effects, and ϵ is a vector of residuals. Variables in $\mathbf{X}$ and $\mathbf{Z}$ may overlap if desired. The LMM can equivalently be represented as a participant-level model:

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i b_i + \epsilon_i.$$

The novelty of this formulation is that it makes clear that not all participants are required to have the same set of measurements for each row of $\mathbf{y}_i$.

The fixed and random effects are assumed to have a multivariate normal joint density

$$\begin{bmatrix} \mathbf{b} \\ \epsilon \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \sigma^2 \right) \tag{2.9}$$

where $\mathbf{G}$ and $\mathbf{R}$ are positive definite matrices and σ^2 is a positive constant. Note that both the random effects b and the residuals ϵ have mean 0. This specification is commonplace for ϵ in OLS regression as the residuals "deviate" from the regression estimate for each participant. The random effects b can be seen similarly. The individual estimates of a random effect are

deviations from a centered normal distribution where the centering occurs on the fixed effect estimate. Using the IQ example from above, if the average baseline level of IQ for the model is 100, then a random effect estimate of 0 indicates the participant has an IQ of 100 (plus ϵ). A random effect estimate of 10 indicates the participant has an IQ of 110 (plus ϵ).

Estimation of the Model Parameters

To find the estimates of β and b, the log likelihood from the joint density (2.9) needs to be maximized. Redefining ϵ as $y - X\beta - Zb$, the joint density of y, b is given by

$$(2\pi\sigma^2)^{-\frac{1}{2}n-\frac{1}{2}q}\left(\det\begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix}^T \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}^{-1} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix}\right\}$$

and the log-likelihood (ignoring the additive constant) is

$$\ell(\beta, \sigma, b) = \frac{1}{2}|R| + \frac{1}{2}|G| + (y - X\beta - Zb)^T R^{-1}(y - X\beta - Zb) + b^T G^{-1} b.$$

As per usual, take the partial derivatives of the log-likelihood function with respect to β and b, set them equal to 0, and solve:

$$\begin{aligned} \frac{\partial \ell}{\partial \beta} &= X^T R^{-1}(y - X\beta - Zb) \\ \frac{\partial \ell}{\partial b} &= Z^T R^{-1}(y - X\beta - Zb) - G^{-1} b \end{aligned}$$

$$\begin{aligned} X^T R^{-1}(y - X\beta - Zb) &= 0 \\ X^T R^{-1} X\beta + X^T R^{-1} Zb &= X^T R^{-1} y \qquad (a) \end{aligned}$$

$$\begin{aligned} Z^T R^{-1}(y - X\beta - Zb) - G^{-1} b &= 0 \\ Z^T R^{-1} X\beta + Z^T R^{-1} Zb + G^{-1} b &= Z^T R^{-1} y \\ Z^T R^{-1} X\beta + (Z^T R^{-1} Z + G^{-1})b &= Z^T R^{-1} y \qquad (b) \end{aligned}$$

where (a) and (b) are known as the Henderson equations (Henderson, 1963):

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & (Z^T R^{-1} Z + G^{-1}) \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix}. \tag{2.10}$$

Notice that the solutions for β and b, respectively $\hat{\beta}$ and $\hat{b}$, contain a "mixture" of the variance of b, $\sigma^2\mathbf{G}$, and the variance of ϵ, $\sigma^2\mathbf{R}$. This mixture is why these models are called mixed models. There are two, separate sources of variance being incorporated into parameter estimation. The variance of the random effects, $\sigma^2\mathbf{G}$, provides information of interest to the estimation of $\mathbf{Y}$. Further, $\hat{b}$ is estimated using partial pooling according to G^{-1}. This pooling of the estimates for the random-effects variables follows from the assumption that b are sampled from a normal distribution. Partial pooling shrinks the estimates of smaller groups, such as those in the tails of a distribution, closer to the mean. Fixed-effects are estimated without any pooling. The implication for the mixed-effects model when contrasted with the OLS regression model is that estimates from a mixed-effects model place greater weight on data highly represented in the sample whereas the estimates from an OLS regression model place equal weight on all data.

An Example of the Utility of LMMs

To demonstrate how this contrast affects estimates in practice, consider this example selected and altered from Robinson (1991).

Example. Nine mice were used in an experiment assessing a pairwise visual discrimination task. After several training sessions, mice were presented with two stimuli representing a reward (S+) and nothing (S-) on a touch screen. Stimuli were randomly presented on either the left or right side of the touch screen over 15 trials with 20 second inter-trial intervals. Each mouse was given between five and seven sessions of 15 trials to learn the task. A successful selection (S+) was rewarded with food whereas an unsuccessful selection (S-) was given a 5 second time-out signal accompanied by house light illumination. After mice had successfully learned the task (deemed by $\geq 80\%$ correct selections on two consecutive sessions), they were placed in a reversal learning condition. This condition reversed such that the successful stimulus was S- whereas the unsuccessful stimulus was S+. The average latency time to respond (in

milliseconds) to each trial during the reversal learning condition was the primary outcome of interest. Table 2.3.1 provides data on each mouse's reaction time.

Table 2.3.1 Reaction Time Latency of 9 mice

Genetic Cohort	Laboratory	Mean Reaction Time (ms)
Syngap1	1	1100
Syngap1	2	1000
Shank2	3	1100
Shank2	2	1000
Shank2	2	1000
Dlgap2	4	1100
Dlgap2	4	1100
Dlgap2	2	1000
Dlgap2	2	1000

The researcher wanted to control for each breed's laboratory as well as the genetic cohort. She decided to use a mixed-effects model assuming $\mathbf{R}$ was the identity matrix (i.e., 1s in the diagonal entries and 0s elsewhere to indicate standard normal, uncorrelated distributions) and $\mathbf{G}$ was one-tenth that size (i.e. $0.1I$). She treated laboratory as a random-effect and genetic cohort as a fixed-effect. Note that the size of $\mathbf{G}$ indicates that she believes the variance of reaction times between laboratories was smaller than between genetic cohorts. Then the parameters of interest are

$$y = \big(1100, 1000, 1100, 1000, 1000, 1100, 1100, 1000, 1000\big)^T$$

$$\beta = \big(\hat{\beta}_{syngap}, \hat{\beta}_{shank}, \hat{\beta}_{dlgap}\big)^T$$

$$b = \big(\hat{b}_1, \hat{b}_2, \hat{b}_3, \hat{b}_4\big)^T.$$

The design matrices for the fixed effects (genetic cohort) $\mathbf{X}$ and random effects (laboratory) $\mathbf{Z}$ were then be set up as

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad Z = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Applying equations (2.10) results in

$$\begin{bmatrix} 2 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 3 & 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 4 & 0 & 0 & 2 & 2 \\ 1 & 0 & 0 & 11 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 11 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 12 & 0 \\ 1 & 2 & 2 & 0 & 0 & 0 & 15 \end{bmatrix} \begin{bmatrix} \hat{\beta}_{syngap} \\ \hat{\beta}_{shank} \\ \hat{\beta}_{dlgap} \\ \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \\ \hat{b}_4 \end{bmatrix} = \begin{bmatrix} 2100 \\ 3100 \\ 4200 \\ 1100 \\ 1100 \\ 2200 \\ 5000 \end{bmatrix}$$

which has solution $\hat{\beta} = (1056.39, 1042.76, 1054.58)^T$ and $\hat{b} = (3.96, 5.20, 7.57, -16.74)^T$. From these results, the researcher may conclude that mutations to *Shank2* is associated with shorter reaction time latency. Further, on the scale of milliseconds, the different laboratories were not much different from one another. If the random effects were treated as fixed effects in this analysis and OLS regression were used (assume laboratory 4 is the referent group for estimability), then the OLS estimates $(X^TX)^{-1}X^Ty$ would be $\hat{\beta}_{ols} = (1000, 1000, 1000, 100, 100, 100)^T$. Note the clear discrepancy between the two models. □

Longitudinal Data Analysis

LMMs are used commonly in longitudinal analyses because the underlying premise of them is that some subset of the regression parameters, the random effects, vary randomly from one individual to another. That is, individuals in the population are assumed to have randomly varying subject-specific trajectories over time (Fitzmaurice, Laird, & Ware, 2011). In this context, the variance components $\mathbf{G}$ and $\mathbf{R}$ can be seen as between-subject and within-subject variance components, respectively. Since subjects have model- and subject-specific estimates

determined through fixed- and random-effects, the LMM provides a useful tool for prediction. That is, conditional on the fixed effects, a participant's individual outcome trajectory can be estimated from that individual's random effects. Longitudinal analyses pose problems for the more standard OLS regression because OLS regression fails to incorporate the appropriate within-subject covariance. Given that repeated measures are typically correlated over time, this results in OLS regression producing biased estimates of the variance (Liang & Zeger, 1993). Therefore, preference should be given to the LMM in the context of longitudinal data analysis.

2.4 Survival Analysis

Censoring

Survival analysis estimates the time from a specified origin until the occurrence of an event. This time is a nonnegative random variable ($T \in [0, \infty)$). However, what separates survival analysis from other statistical methods is the presence of censoring. Censoring occurs when the event of interest and its time are not observed. In these cases, a censored observation contains only partial information about the random variable of interest. Therefore, censored observations need to be accounted in the analysis of the time-to-event. Censoring is defined based on the interval of time not observed. In general, there is *left censoring*, *interval censoring*, and *right censoring*. Left censoring is defined by the event of interest occurring prior to the first observation (e.g., enrollment into a study). Interval censoring is defined by the event of interest occurring between two observations. Right censoring is defined by the event of interest not occurring during the period of observation (i.e., the event of interest occurs after the study period). Figure 2.4.1 provides examples of the different types of censoring.

In this illustration, ID 1 was fully observed until failure where 'failure' refers to the event transpiring within survival analysis literature. ID 2 was observed during the whole period of the study, but the event occurred prior to enrollment. Therefore, ID 2 was left censored. ID 3 was observed from years 0 to 3 and from 7 to 10. However, the event occurred during an unobserved period of time. ID 3 has interval censoring. ID 4 was observed until 9 years into

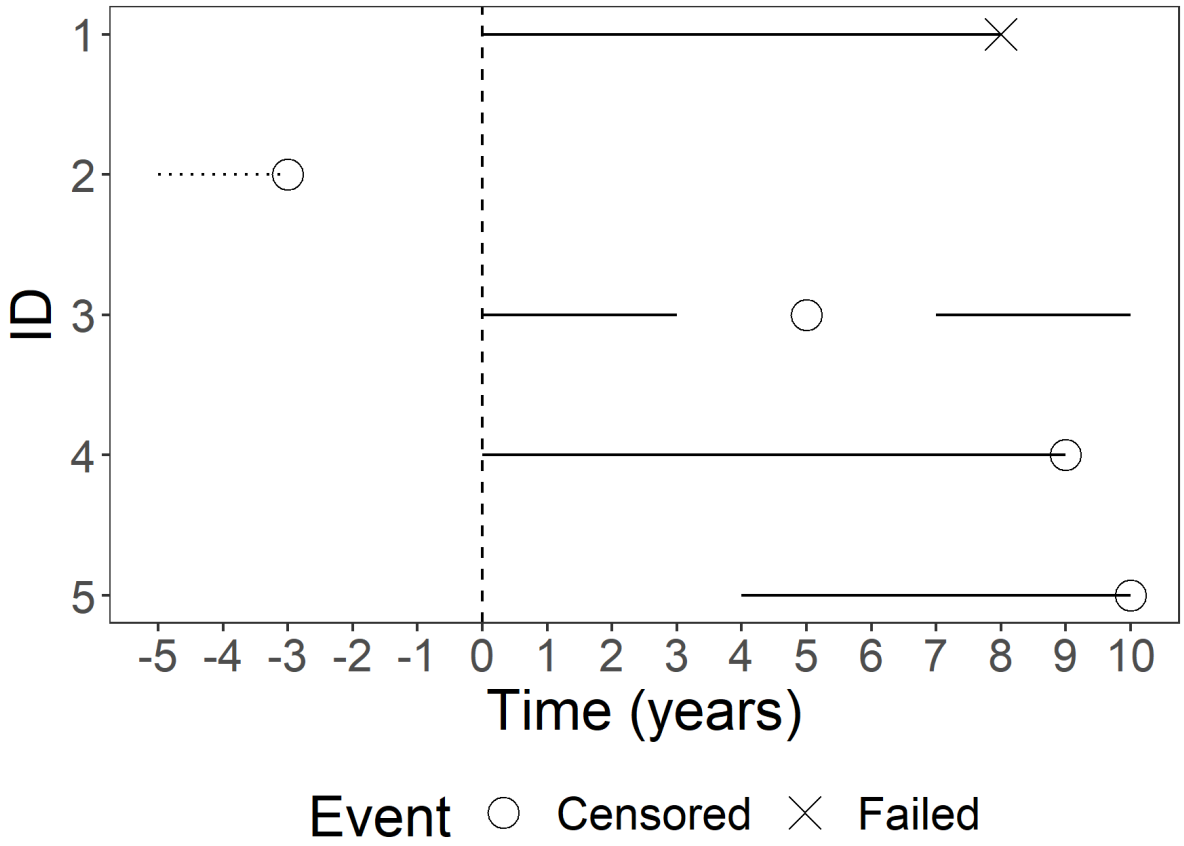

Figure 2.4.1 Examples of Left, Right, and Interval Censoring

the study without the event of interest occurring. ID 5 was observed until the end of the study without the event occurring. Therefore, both ID 4 and 5 were right censored. The censoring time for ID 4 was 9 years whereas the censoring time for ID 5 was 6 years since this participant was enrolled in year 4.

Out of all forms of censoring, right censoring occurs most often in practice and as such receives the most focus within texts on survival analysis. Common reasons for right censoring are loss to follow up, drop out, and the termination of a study prior to event occurrence. Left censoring is not uncommon in observational studies and certain clinical studies, but often participants who have some left censoring may be excluded (i.e., left truncated) from the study design. Left censoring was found among many participants in the T4+ and AIDS development studies from Chapter 1, where several participants in the ZDT trials had been infected with HIV for an unknown amount of time prior to enrollment. In theory, interval censoring is often treated as the general case of censoring where left and right censoring are considered special cases (Zhang & Sun, 2010). In practice, interval censoring is more likely when observation periods are farther apart. One consideration with interval censoring is whether the most recent observation should be treated as the observed time. In observational or clinical studies, almost certainly the exact timing of an event will not occur during the study's measurement periods, and there will be some measurement error in participant recall. However, this does not necessitate treating all observations as interval censored. Rather, this can depend on the hypothesis being tested.

These three types of censoring are sometimes called *random censoring*. In this manuscript, all censoring is considered random. For the sake of posterity, nonrandom censoring comprises what Miller (1981) designates "Type I" and "Type II" censoring. Type I censoring occurs when some fixed value of time is assigned as the fixed censoring time. All observations that do not fail within that time period are considered censored. This type of censoring occurs in engineering applications where a series of products, such as light bulbs, are used until burn out. These experiments typically have a predetermined time period established from the expected value of burnouts among the product. Type II censoring occurs when some fixed number of failures

occur. After r number of failures occur, all other observations are considered censored. Using the bulb example, an experiment that observes bulb burn out until 50% of the sample fails will result in the other 50% of the bulbs having Type II censoring.

Ignoring censoring produces bias in survival probability estimation, hence why this topic receives special attention. This bias can either be an overestimation or underestimation of the survival probability, so a post-hoc naive correction likely would not resolve the issues posed by ignoring the censoring. Further, mishandling the censoring of data can lead to biased estimates. Censoring is assumed to be *noninformative* or random. That is, the time to failure and the time to censoring are independent. Censoring that is not noninformative is described as informative or nonrandom censoring, where time to censoring and time to failure are correlated. Informative censoring may occur in a randomized clinical trial where a disproportionate number of treatment group participants are censored. The treatment may be very effective, leading the treatment group to be loss to followup due to no longer needing treatment. The treatment may also be iatrogenic causing early death. Both of these scenarios would cause bias in the survival probability as the true failure time is masked by the dependent censoring time. Informative censoring requires specialized methods, directly related to the censoring or missingness mechanism. A discussion on this topic can be found in Diggle and Kenward (1994).

The Functions of Survival Analysis

Routine summary statistics, such as the mean and standard deviation, are likely biased in the presence of censored data. Therefore, survival data are estimated with a set of distribution functions, from which modified summary statistics may be estimated. The two primary functions in survival analysis are the survival function $S(t)$ and the hazard function $\lambda(t)$. However, first consider two distributions that are ubiquitous throughout statistics: the cumulative distribution function, $F(t)$, and the probability density function, $f(t)$.

In the context of survival probabilities, $F(t)$ is defined as the probability of failure up to some nonnegative value of time T. That is,

$$F(t) = P(T \leq t), \quad t \geq 0.$$

By definition, it is assumed $F(0) = 0$ and $F(\infty) = 1$. Since $F(t)$ is a probability, it will always be bound between 0 and 1. As is the case in all of statistics, the probability density function is just the derivative of the cumulative distribution function:

$$f(t) = \frac{dF(t)}{dt},$$
$$F(t) = \int_0^t f(u)du$$

The value $f(t)$ can be interpreted as a probability of failure "at" some time t, so the probability of failure up to time t, $F(t)$, is the sum of these individual probabilities from the beginning of the study $t = 0$ until t. To prevent the belaboring of certain technical points, consider this aforementioned description of $f(t)$ as a heuristic one.

The survival probability, $S(t)$, is the inverse of the failure probability.

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

This definition naturally implies

$$f(t) = -\frac{dS(t)}{dt},$$
$$S(t) = \int_t^\infty f(u)du$$

where these formulas can be determined by substituting $1 - S(t)$ into the equations for $F(t)$. From this definition of the survival probability, both the mean and median survival times

are respectively estimated by $E(T) = \mu = \int_0^\infty S(t)dt$ and finding the smallest t such that $S(t) \leq 0.5$.

The hazard function defines the rate of failure "at" time t given the individual has not failed up to time t. Once again, consider this a heuristic explanation in order to minimize the discussion of technical details. This definition can be akin to a conditional probability as in §2.2:

$$\lambda(t) = \frac{f(t)}{S(t)}.$$

Using previous definitions and formula, $\lambda(t)$ can be determined in many different ways.

$$\begin{aligned}\lambda(t) &= \frac{f(t)}{S(t)} \\ &= -\frac{S'(t)}{S(t)} \\ &= -\frac{d \log S(t)}{S(t)}\end{aligned}$$

In practice, the log hazard function instead is estimated $\alpha(t) = \log \lambda(t)$. As with the probability density function, the cumulative hazard function–the total hazard up to time t–can be estimated by integrating from $t = 0$ to t:

$$\begin{aligned}\Lambda(t) &= \int_0^t \lambda(u)du = -\log S(t), \\ S(t) &= \exp\{-\Lambda(t)\} = \exp\{-\int_0^t \lambda(u)du\}\end{aligned}$$

In summary, the survival probability equations (2.11) are provided below.

$$
\begin{aligned}
F(t) &= \int_0^t f(u)du = 1 - S(t), \\
S(t) &= 1 - F(t) = \int_t^\infty f(u)du = \exp\{-\Lambda(t)\} = \exp\{-\int_0^t \lambda(u)du\}, \\
\lambda(t) &= \frac{f(t)}{S(t)} = -\frac{d\log S(t)}{dt}, \\
\Lambda(t) &= \int_0^t \lambda(u)du = -\log S(t), \\
f(t) &= \frac{dF(t)}{dt} = -\frac{dS(t)}{dt} = \lambda(t)S(t).
\end{aligned}
\tag{2.11}
$$

Parametric, Nonparametric, and Semiparametric Estimation

Statistical models fall into one of three categories based on the assumptions made for estimation: parametric models, nonparametric models, and semiparametric models. Parametric models assume that the distribution of the random variable is determined by a set of parameters. Once these parameters are known or estimated, the entire model is known. For example, a normal distribution is determined by two parameters, its mean μ and variance σ^2. Ordinary least squares regression assumes a linear relationship between the β vector of regression coefficients and the normally distributed outcome. Statistical models that do not make any assumptions are nonparametric models. In a sense, nonparametric models are entirely data-driven. Rather than make assumptions about the distribution of the data, a nonparametric model lets the data dictate how they are distributed. A prime example of a nonparametric model is k-means clustering, which clusters data around an adaptively determined number of centroids (Lloyd, 1982). Semiparametric models lie somewhere in between these two extremes. Some assumptions are made about the distribution of the data whereas some properties are data-driven. The Gaussian mixture model in §2.2 is semiparametric, as it assumes normal parameters to describe the comprising distributions but also lets the data decide the proportion of members who belong to one distribution versus another.

The power of the generalized linear model resulted in parametric models dominating the other two categories in most regression settings. However, survival analysis commonly employs all three categories. Their use in survival analysis is so ubiquitous that survival models are the common examples of nonparametric and semiparametric methods used in introductory papers on the topics (Dudley, Wickham, & Coombs, 2016; Kosorok, 2009). Some discussion on all three methods is provided here.

Table 2.4.1 lists common parametric models of survival analysis. Note that $\Gamma(\cdot)$ refers to the gamma function where $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$ and $\Phi(\cdot)$ refers to the standard normal cumulative distribution function. The exponential distribution is the simplest of these functions in part because $\lambda(t)$ is constant. This property of the exponential distribution's constant hazard is the memoryless property–the value of the hazard is not contingent on time, and past failure does not influence current or future failure. The Weibull distribution is the rough equivalent

Table 2.4.1 Common Parametric Models in Survival Analysis

Distribution	$f(t)$	$S(t)$	$\lambda(t)$	$E(T)$
Exponential	$\lambda e^{-\lambda t}$	$e^{-\lambda t}$	λ	$\frac{1}{\lambda}$
Weibull	$\psi\lambda t^{\psi-1}e^{-\lambda t^\psi}$	$e^{-\lambda t^\psi}$	$\psi\lambda t^{\psi-1}$	$\frac{\Gamma(1+\frac{1}{\psi})}{\lambda^{1/\psi}}$
Log-normal	$\frac{1}{\sqrt{2\pi}\sigma t}e^{-\frac{(\log t-\mu)^2}{2\sigma^2}}$	$1-\Phi(\frac{1}{\sigma}\{\log t-\mu\})$	$\frac{f(t)}{S(t)}$	$e^{\mu+\frac{\sigma^2}{2}}$

of the normal distribution in survival analysis. Its survival and hazard functions change with respect to a scaling parameter ψ. Note that if $\psi = 1$, then the Weibull distribution reduces to the exponential distribution. A positive ψ produces an increasing hazard, and a negative ψ produces a decreasing hazard. The log-normal distribution is the most complex of these distributions. Its advantage over the previous two distributions lies in the hump shape of its hazard function. That is, the hazard increases for a period of time and then decreases after reaching its peak. Figure 2.4.2 illustrates various hazard functions for these three distributions.

Suppose a sample of n individuals have random survival times $T_1, T_2, \ldots, T_n$. Not all of these times may be observed due to censoring, so further suppose these individuals also have random censoring times $C_1, C_2, \ldots, C_n$. Then each participant has a pair of random

variables $(T_1, C_1), (T_2, C_2), \ldots, (T_n, C_n)$. In cases where survival times are observed, consider the censoring time as "potential" censoring time. For example, if the first participant had a survival time of 3 years and was not censored, then $T_1 = 3$. In this case, C_1 could potentially be 4 years, 5.7 years, or some other value greater than 3. Therefore, while (T_i, C_i) are the random variables of interest, they are not directly observed. Rather, a time to event X_i and the censoring status Δ_i of that event are both observed. Therefore, the data collected are $(X_1, \Delta_1), (X_2, \Delta_2), \ldots, (X_n, \Delta_n)$. Formally, the observed data are defined

$$X_i = \min(T_i, C_i) \qquad \text{(time to observed event)}$$

$$\Delta_i = I(T_i \leq C_i) = \begin{cases} 1 & \text{if } T_i \leq C_i \text{ (failure time observed)} \\ 0 & \text{if } T_i > C_i \text{ (censoring time observed).} \end{cases}$$

Since the inference is made on (T_i, C_i), the likelihood function for right censored parametric survival functions will incorporate parameters θ for T_i and ϕ for C_i. Let x_i be the observed time to event and $\delta_i = \{1, 0\}$ depending on whether the observed time to event was failure (= 1) or censoring (= 0). The likelihood function for θ and ϕ needs to include the density function $f(t)$ and survival function $S(t)$ for θ as they represent the probability of failure at a given time t and the probability of survival beyond time t (or the probability of being "potentially" censored). Equivalent functions need to be incorporated for ϕ, which will be called $g(t)$ and $H(t)$, respectively. Therefore, the likelihood function is

$$L(\theta, \phi | x_i, \delta_i) = \prod_{i=1}^{n} \{[f(x_i|\theta)]^{\delta_i}[S(x_i|\theta)]^{1-\delta_i}\}\{[g(x_i|\phi)]^{1-\delta_i}[H(x_i|\phi)]^{\delta_i}\}. \tag{2.12}$$

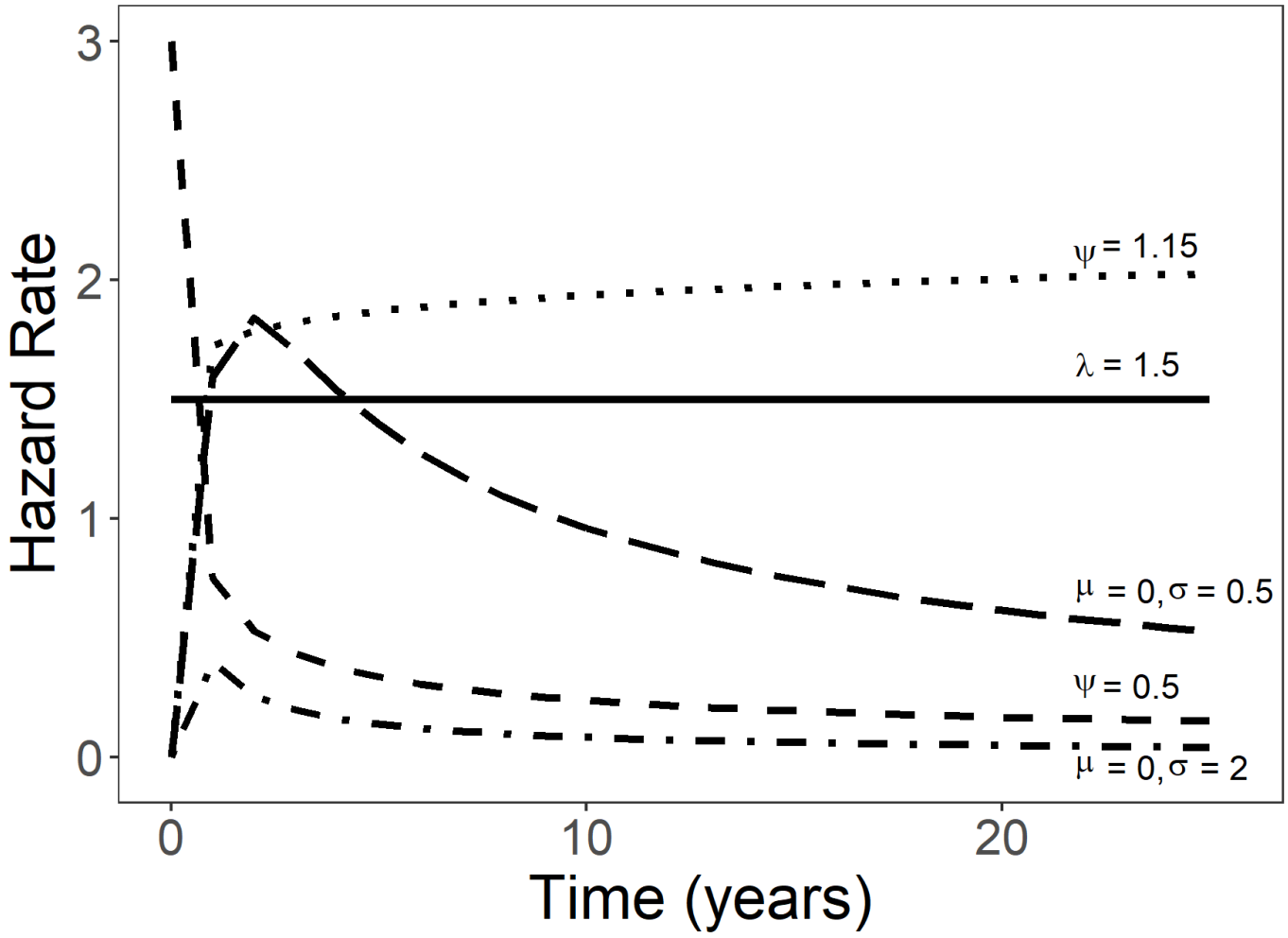

Figure 2.4.2 Hazard Functions of exponential distribution with $\lambda = 1.5$ (solid), Weibull distribution with $\psi = 0.5$ (dashed), Weibull distribution with $\psi = 1.15$ (dotted), Lognormal distribution with $\mu = 0, \sigma = 0.5$ (long dashed), and Lognormal distribution with $\mu = 0, \sigma = 2$ (dot dashed)

Let ν represent the number of participants where failures were observed. Note what happens in equation (2.12) when times X_i are observed failures (i.e. $\delta_i = 1$):

$$
\begin{aligned}
L(\theta, \phi | x_i, \delta_i = 1) &= \prod_{i=1}^{\nu} \{[f(x_i|\theta)]^1 [S(x_i|\theta)]^{1-1}\} \{[g(x_i|\phi)]^{1-1} [H(x_i|\phi)]^1\} \\
L(\theta, \phi | x_i, \delta_i = 1) &= \prod_{i=1}^{\nu} \{[f(x_i|\theta)]^1 [S(x_i|\theta)]^0\} \{[g(x_i|\phi)]^0 [H(x_i|\phi)]^1\} \\
L(\theta, \phi | x_i, \delta_i = 1) &= \prod_{i=1}^{\nu} \{[f(x_i|\theta)] \times [1]\} \{[1] \times [H(x_i|\phi)]\} \\
L(\theta, \phi | x_i, \delta_i = 1) &= \prod_{i=1}^{\nu} [f(x_i|\theta)][H(x_i|\phi)].
\end{aligned}
$$

The observed failures contribute to the probabilities of failing at time x_i and "potentially" failing from time x_i onward. Under noninformative censoring where T_i is independent from C_i, then the parameters refering to C_i can be ignored. Therefore, equation (2.12) becomes $L(\theta | x_i, \delta_i) = \prod_{i=1}^{n} \{[f(x_i|\theta)]^{\delta_i} [S(x_i|\theta)]^{1-\delta_i}\}$. Using the definition of $\lambda(t)$ from equation (2.11), $\lambda(t) = \frac{f(t)}{S(t)} \leftrightarrow f(t) = \lambda(t) S(t)$. This substitution for $f(t)$ simplifies (2.12) further:

$$
\begin{aligned}
L(\theta | x_i, \delta_i) &= \prod_{i=1}^{n} [f(x_i|\theta)]^{\delta_i} [S(x_i|\theta)]^{1-\delta_i} \\
L(\theta | x_i, \delta_i) &= \prod_{i=1}^{n} [\lambda(x_i|\theta) S(x_i|\theta)]^{\delta_i} [S(x_i|\theta)]^{1-\delta_i} \\
L(\theta | x_i, \delta_i) &= \prod_{i=1}^{n} [\lambda(x_i|\theta)]^{\delta_i} [S(x_i|\theta)]^{\delta_i + 1 - \delta_i} \\
L(\theta | x_i, \delta_i) &= \prod_{i=1}^{n} [\lambda(x_i|\theta)]^{\delta_i} [S(x_i|\theta)]^1 \\
L(\theta | x_i, \delta_i) &= \prod_{i=1}^{n} [\lambda(x_i|\theta)]^{\delta_i} [S(x_i|\theta)] \qquad (2.13)
\end{aligned}
$$

indicating that for the parametric case, the likelihood function can be constructed from the observed data (X_i, Δ_i), the hazard function $\lambda(x_i)$, and the survival function $S(x_i)$.

The primary nonparametric methods used in survival analysis are the Kaplan-Meier estimator of the survival function (Kaplan & Meier, 1958) and the Nelson-Aalen estimator of the cumulative hazard function (Nelson, 1969; Aalen, 1976). Since nonparametric methods make no assumptions about the distribution of (T_i, C_i), the likelihood function given in (2.13) cannot be used to estimate the time to event. Rather than a likelihood approach, these two estimators use a product limit method to estimate the survival function and cumulative hazard function. Consider the whole interval of time where $\nu_1 < \nu_2 < \ldots \nu_J$ are the ordered observed times. Let all X_i fall into an interval $[\nu_j, \nu_{j+1}]$. Therefore, the observed times $x_1, x_2, \ldots, x_n$ are now represented as intervals $[\nu_j, \nu_{j+1}]$. Let $d(x)$ be the total number of deaths within any interval $[\nu_j, \nu_{j+1}]$ and let $n(x)$ represent the number of individuals still at risk in that interval. Then,

$$\hat{S}_{KM} = \prod_{j=1}^{J}(1 - \frac{d(x_j)}{n(x_j)})$$
$$\hat{\Lambda}_{NA} = \sum_{x_j \leq x} \frac{d(x_j)}{n(x_j)},$$

where $\hat{S}_{KM}$ is the Kaplan-Meier estimator and $\hat{\Lambda}_{NA}$ is the Nelson-Aalen estimator. These estimators produce step functions of the survival and hazard functions, respectively, and are useful for illustrating life table estimates of survival data. Consider Table 2.4.2 and Figure 2.4.3.

Table 2.4.2 Life Table of Simulated Survival Data

Time	Failures	Censors	$\hat{S}_{KM}$	$\hat{\Lambda}_{NA}$
0	0	0	1.00	0.00
9	1	0	0.917	0.08
17	1	0	0.833	0.174
22	0	1	0.833	0.174
30	1	1	0.741	0.285
41	2	0	0.529	0.571
60	1	0	0.423	0.771
77	0	1	0.423	0.771
80	0	1	0.423	0.771
106	1	0	0.212	1.271
150	1	0	0	2.271

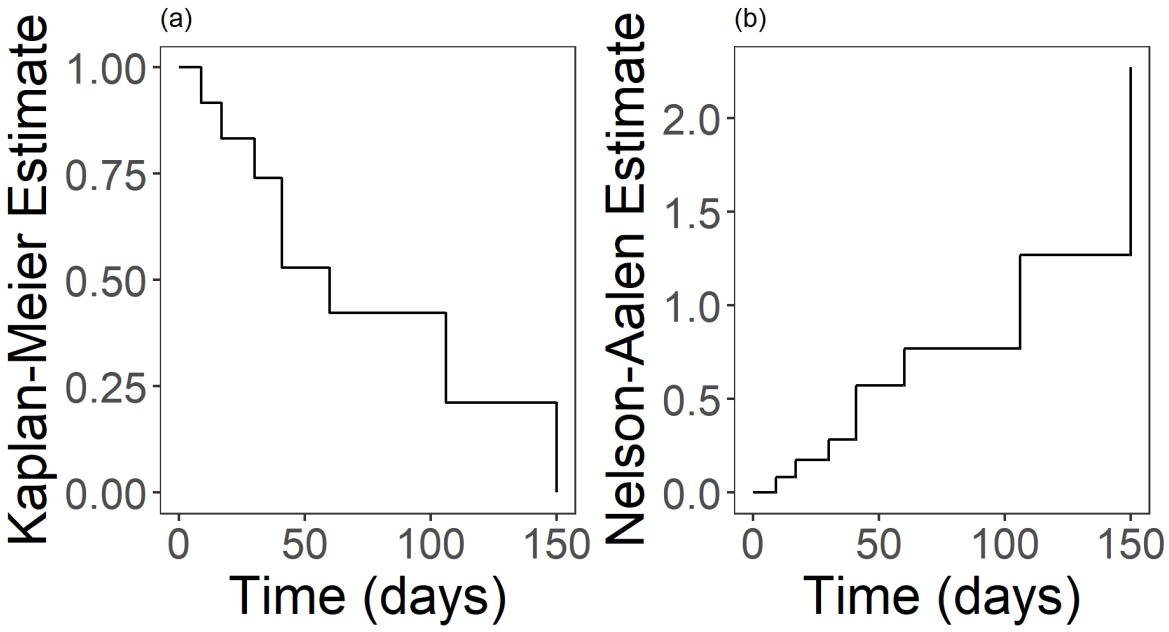

Figure 2.4.3 (a) Kaplan-Meier Survival Estimate (b) Nelson-Aalen Cumulative Hazard Estimate

The most common semiparametric method within survival analysis, and perhaps the most commonly used model in survival analysis, is the Cox proportional hazards model (Cox, 1972). The central idea of this model is that the hazard function $\lambda(t)$ can be estimated using a data-driven approach. Let $\mathbf{Z}$ represent covariates. Recall that $\lambda(t) = \frac{f(t)}{S(t)}$. The hazard function in a Cox proportional hazard function is defined by $\lambda(t|z) = \frac{f(t|z)}{S(t|z)} = \lambda_0(t)e^{Z^T\beta}$, where β are regression coefficients of $\mathbf{Z}$ and $\lambda_0(t)$ is some baseline hazard function. The baseline hazard function is arbitrary, and the focus is on estimation of the change in the hazard relative to the change in the covariates. The unit $e^{Z^T\beta}$ is treated as a separate hazard function that is factored out of the baseline hazard in that $e^{Z^T\beta} := \lambda(Z^T\beta)$. The central assumption for the Cox proportional hazards model is that the data-driven hazard function $\lambda(t|z)$ is proportional to a common baseline hazard $\lambda_0(t)$ and an effect on the covariates $\mathbf{Z}$. Note that the effect of time is separable from the effects of the covariates, where t only affects the hazard function through the baseline hazard itself. This separability is called the proportional hazards assumption. Let the covariate of interest be a dichotomous variable, Z, where $Z = \{1, 0\}$. The hazard ratio

comparing participants with $Z = 1$ to those with $Z = 0$ is estimated as

$$\begin{aligned} HR &= \frac{\lambda_0(t)e^{\beta_0+\beta_1(Z=1)}}{\lambda_0(t)e^{\beta_0+\beta_1(Z=0)}} \\ &= e^{\beta_1(1-0)} \\ &= e^{\beta_1}. \end{aligned}$$

The change in $\lambda(t|z)$ is simply a result of a change in Z.

The construction of the likelihood function for this method requires some nuance as the method uses some parametric assumptions and nonparametric ones. Consider the pair of random variables (T_i, C_i) as detailed in the parametric section. The random variable T_i is the primary interest of survival analysis, and C_i must be accounted for due to its presence in estimating T_i. In general, C_i is considered a *nuisance* parameter, which means exactly how it sounds: there is little interest in inference about C_i, but it has to be dealt with to provide inference for T_i. Order the data into intervals similarly to the nonparametric product-limit estimation methods. Then the likelihood function is given by

$$\begin{aligned} L(\mathbf{T}, \mathbf{C}|\beta) &= \prod_{i=1}^{m} p(T_i, C_i|(T_1, C_1), (T_2, C_2), \ldots, (T_{i-1}, C_{i-1}), \beta) \\ &= \prod_{i=1}^{m} p(C_i|(T_1, C_1), (T_2, C_2), \ldots, (T_{i-1}, C_{i-1}), \beta) \\ &\times \prod_{i=1}^{m} p(T_i|(T_1, C_1), (T_2, C_2), \ldots, (T_{i-1}, C_{i-1}), \beta). \end{aligned} \tag{2.14}$$

Cox (1975) defines the second factor (2.14) as the partial likelihood and proves that most of the information about β in a Cox proportional hazard model is found in this factor. Therefore, the first factor can be safely ignored given a large enough sample size. Miller (1981) demonstrates that given observed values (X_i, Δ_i), this partial likelihood is $L(\beta|x_i, \delta_i) =$

$\prod_{i=1}^{n}[\lambda(x_i|z_i,\beta)]^{\delta_i}[S(x_i|z_i,\beta)]$. Note the parallel between this partial likelihood and the likelihood given in equation (2.13).

2.5 B-splines

In mathematics, splines are functions defined by piecewise polynomial functions typically intended for the interpolation of points. If a function is unknown, splines provide an approximation of that function using polynomials. Since there are many ways to approximate functions using piecewise polynomial functions, B-splines offer a unique, piecewise polynomial representation of a function given a set of specified continuity conditions. While there are many ways to formulate B-splines, this manuscript uses the recursive definition specified in de Boor (1986) and Kincaid & Cheney (1992). A B-spline is uniquely identified by its continuity conditions: its knot sequence and its degree. The knot sequence $\mathbf{t} := (t_1, t_2, \ldots, t_n)$ form the partitions or continuity points of the B-spline. Essentially, these are the locations where the points are "connected". The degree specifies the degree of the piecewise polynomials. For example, degree $k = 3$ indicates the B-spline will be a cubic function (i.e., the biggest term will be some function of t^3).

All B-splines will have a degree = 0 component where

$$B_j^0(x) = \begin{cases} 1 & \text{if } t_j \leq x < t_{j+1} \\ 0 & \text{otherwise.} \end{cases}$$

This B-spline behaves similarly to an indicator function that determines whether a point lies within a knot sequence. Then, all other B-splines are defined recursively where

$$\begin{aligned} B_j^{(k)}(x) &= \omega_j^{(k)} B_j^{(k-1)}(x) + (1 - \omega_{j+1})^{(k)} B_{j+1}^{(k-1)}(x) \\ &= \left(\frac{x - t_j}{t_{j+k} - t_j}\right) B_j^{(k-1)}(x) + \left(\frac{t_{j+k+1} - x}{t_{j+k+1} - t_{j+1}}\right) B_{j+1}^{(k-1)}(x). \end{aligned}$$

B-splines are restricted to sum to 1 within any degree (e.g., $\sum B_j^2 = 1$) and each B-spline must be nonnegative.

Example. Define a knot sequence $\mathbf{t} := (1, 2, 4, 6, 8)$. A B-spline of degree $k = 2$ is constructed as:

$$B_j^0 = \begin{cases} 1 & \text{if } t_j \leq x < t_{j+1} \\ 0 & \text{otherwise.} \end{cases}$$

indicating $B_1^0 = 1$ when $x \in [1, 2)$, $B_2^0 = 1$ when $x \in [2, 4)$, $B_3^0 = 1$ when $x \in [4, 6)$, $B_4^0 = 1$ when $x \in [6, 8)$, and 0 otherwise. Next,

$$B_j^1 = \left(\frac{x - t_j}{t_{j+1} - t_j} \right) B_j^0(x) + \left(\frac{t_{j+2} - x}{t_{j+2} - t_{j+1}} \right) B_{j+1}^0(x)$$

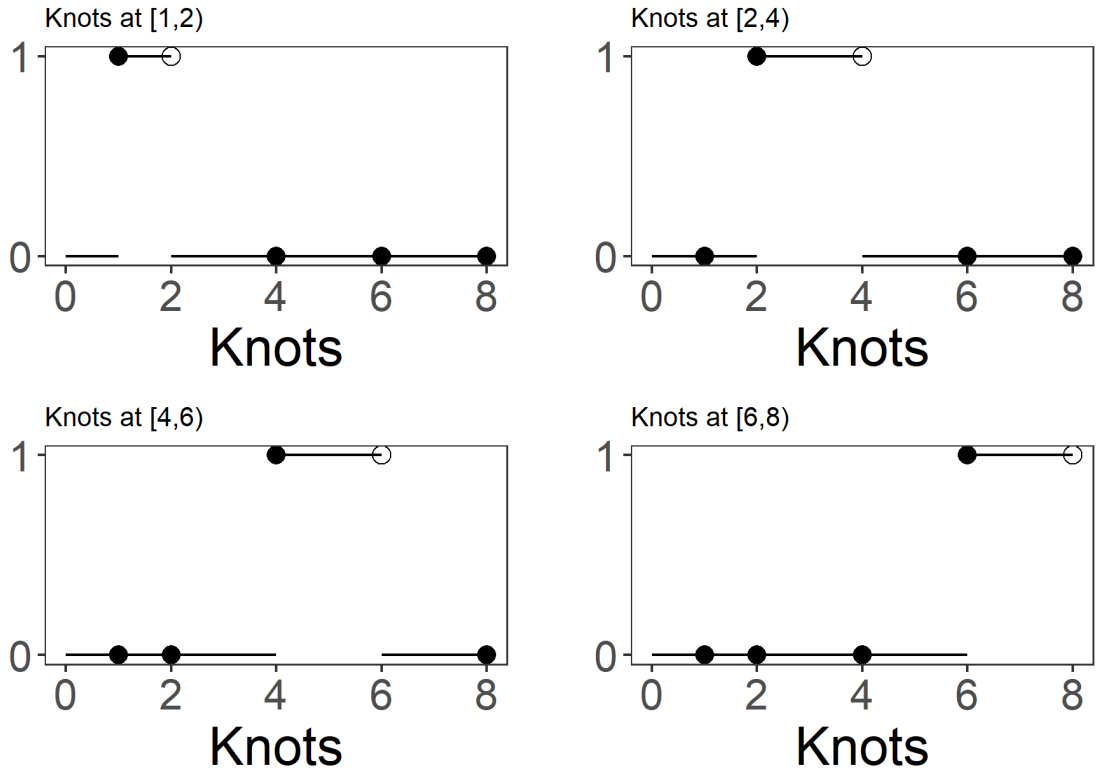

Figure 2.5.1 B-Splines for Degree 0 with knot sequence (1, 2, 4, 6, 8)

which implies

$$\rightarrow B_1^1 = \begin{cases} x-1 & \text{if } 1 \leq x < 2 \\ 2 - \frac{1}{2}x & \text{if } 2 \leq x < 4 \\ 0 & \text{otherwise} \end{cases}$$

$$\rightarrow B_2^1 = \begin{cases} \frac{1}{2}x - 1 & \text{if } 2 \leq x < 4 \\ 3 - \frac{1}{2}x & \text{if } 4 \leq x < 6 \\ 0 & \text{otherwise} \end{cases}$$

$$\rightarrow B_3^1 = \begin{cases} \frac{1}{2}x - 2 & \text{if } 4 \leq x < 6 \\ 4 - \frac{1}{2}x & \text{if } 6 \leq x < 8 \\ 0 & \text{otherwise} \end{cases}$$

because, for example, $B_1^1 = \left(\frac{x-t_2}{t_3-t_2}\right) I[x \in [1,2)] + \left(\frac{t_3-x}{t_3-t_2}\right) I[x \in [2,4)] = \left(\frac{x-1}{2-1}\right) I[x \in [1,2)] + \left(\frac{4-x}{4-2}\right) I[x \in [2,4)] = (x-1)I[x \in [1,2)] + (2 - \frac{1}{2}x)I[x \in [2,4)]$. Finally, for degree $k = 2$,

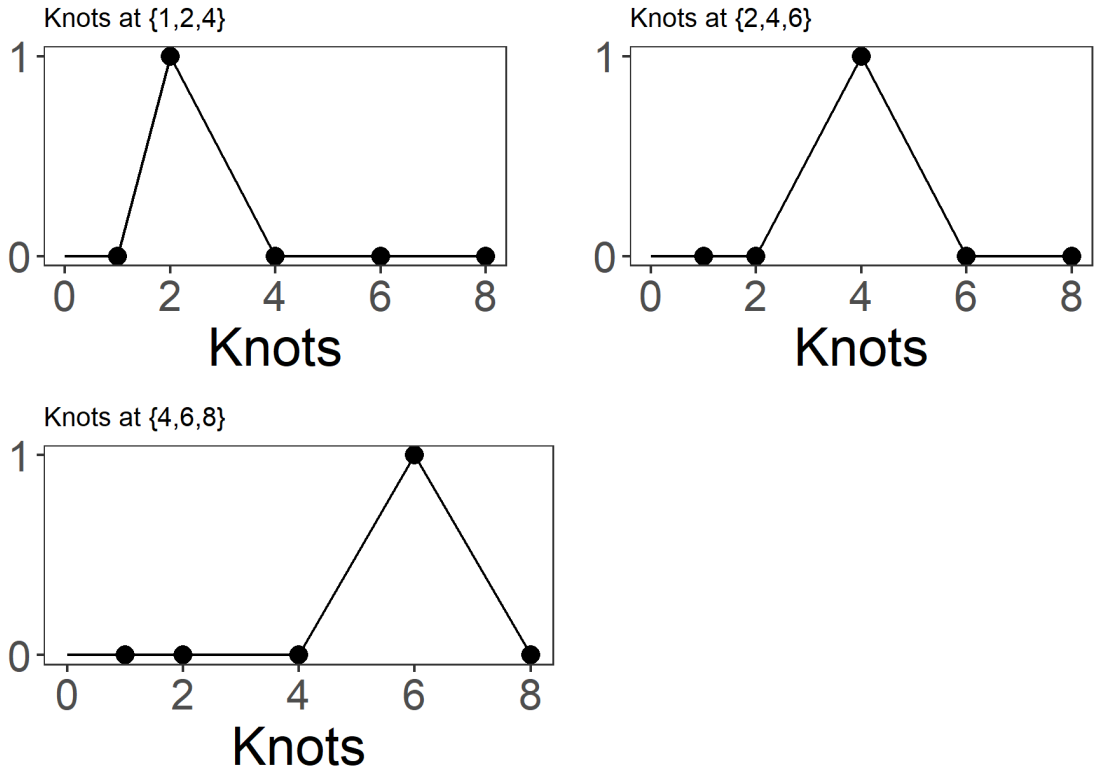

Figure 2.5.2 B-Splines for Degree 1 with knot sequence (1, 2, 4, 6, 8)

$$\rightarrow B_1^2 = \begin{cases} \frac{1}{3}(x-1)^2 & \text{if } 1 \leq x < 2 \\ -\frac{7}{24}x^2 + x - \frac{13}{6} & \text{if } 2 \leq x < 4 \\ \frac{1}{2}x^2 - 6x + \frac{9}{2} & \text{if } 4 \leq x < 6 \\ 0 & \text{otherwise} \end{cases}$$

$$\rightarrow B_2^2 = \begin{cases} \frac{1}{6}(x^2 - \frac{5}{2}x + 1) & \text{if } 2 \leq x < 4 \\ -\frac{7}{24}x^2 = \frac{9}{2}x - 10 & \text{if } 4 \leq x < 6 \\ -\frac{1}{8}x^2 - 2x + 8 & \text{if } 6 \leq x < 8 \\ 0 & \text{otherwise} \end{cases}$$

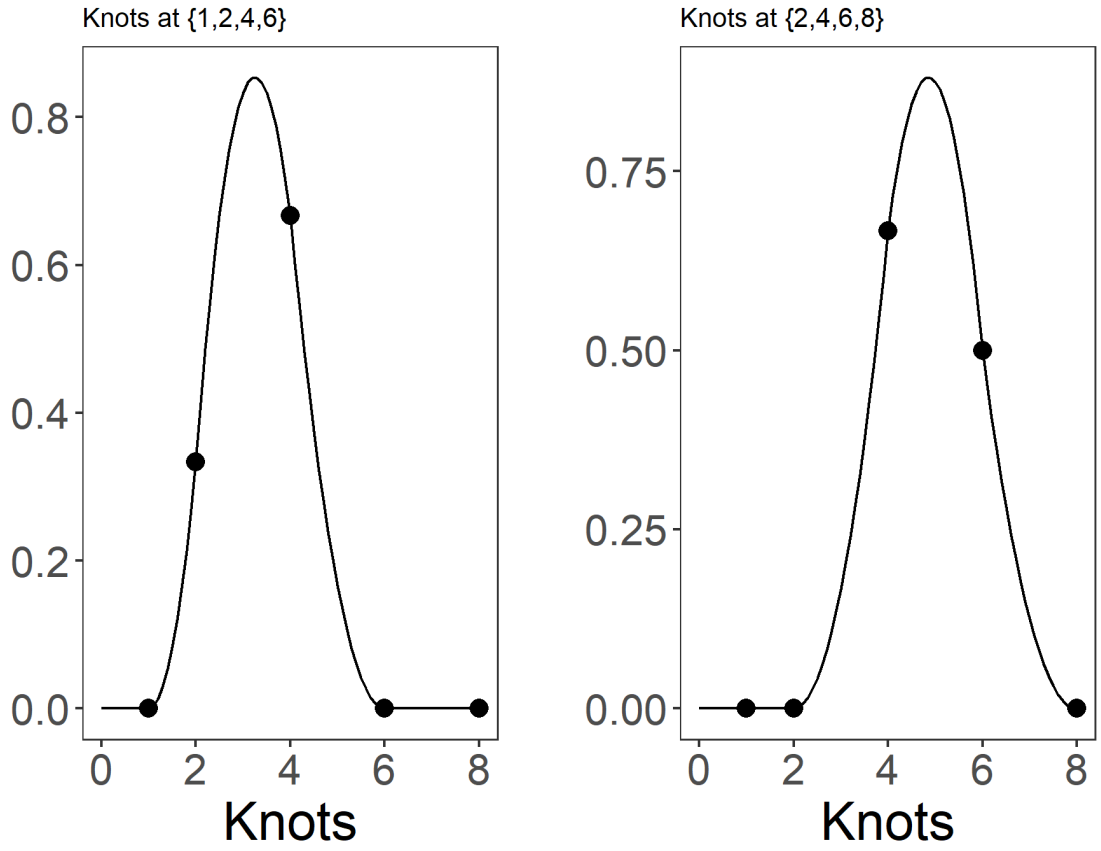

Figure 2.5.3 B-Splines for Degree 2 with knot sequence (1, 2, 4, 6, 8)

Notice how the number of knots needed for each B-spline increases as the degree increases. Therefore, consideration is needed when determining the knot sequence with respect to the degree of the B-spline.

2.6 Tensor Products

Given matrices $\mathbf{v}$ and $\mathbf{w}$, and assuming the matrices are conformable, the *tensor product* or *outer product* is defined as $\mathbf{v} \otimes \mathbf{w} = \mathbf{v}\mathbf{w}^T$. Let $\mathbf{v} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ and $\mathbf{w} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$, then the tensor product

is $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \begin{bmatrix} 4 & 5 \end{bmatrix} = \begin{bmatrix} 1 \times 4 & 1 \times 5 \\ 2 \times 4 & 2 \times 5 \\ 3 \times 4 & 3 \times 5 \end{bmatrix} = \begin{bmatrix} 4 & 5 \\ 8 & 10 \\ 12 & 15 \end{bmatrix}$. The tensor product effectively produces extra dimensions from lower dimensional spaces based on the permutations of each matrix.

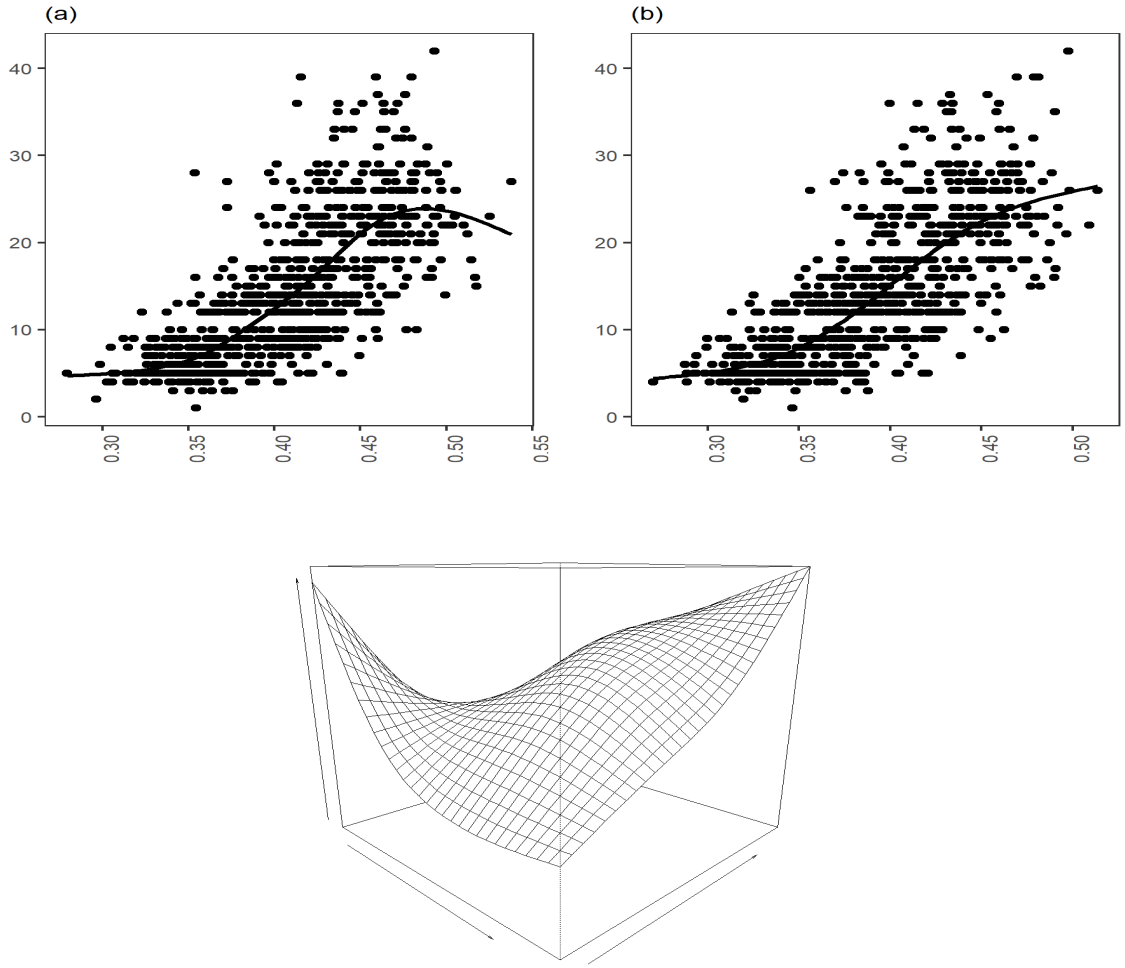

Figure 2.6.1 Tensor Product Visualization

In additive models (Friedman & Silverman, 1989) the tensor product can be used to model interactions of the form $y = f(x,t) + \epsilon$ by decomposing the function into two-dimensional basis spaces $f(x,t) = f_x(x)f_t(t)$ and finding the tensor products of these two spaces. This approach has been used in generalized additive models (Hastie & Tibshirani, 1987) and hazard

regression (Kooperberg, Stone, & Truong, 1995a). Figure 2.6.1 illustrates the tensor product between two variables in (a) and (b).

2.7 Stepwise Models and Recursive Partitioning

Stepwise regression is an automated variable selection method used in generalized linear models with particular occurrence in ordinary least squares (OLS) regression model building. This method typically has three components: (1) an inclusion/exclusion rule, (2) an assessment method, and (3) a stopping rule. The inclusion/exclusion rule determines whether a variable should be included in the model based on some statistic. The assessment method calculates the inclusion statistic. The stopping rule determines when the selection process should stop. Stepwise procedures have forward processes (variables are added individually from a null model until the "best" model is determined), a backward process (variables are removed from the saturated model individually until the "best" model is determined), or both. Inclusion/exclusion rules typically used include partial-F tests or p-values, where inclusion in the model could be set to the threshold of having a p-value < 0.15. However, information-based methods such as minimizing AIC (Aikake, 1974) has also been used in these methods. Given two variables, X_1 and X_2, an OLS regression with forward-backward selection algorithm could look like this:

Setup : Let $\alpha_p = 0.15$. Stop when all $2^k - 1 = 2^2 - 1 = 3$ models tested.

Step 1 : Add X_1 to the null OLS model.

Step 2 : Test p-value for X_1.

Step 3 : Is p-value for $X_1 < \alpha_p$?

Step3a : If yes, include.

Step3b : If no, exclude.

Step 4 : Add X_2 to the OLS model.

Step 5 : Test p-value for X_2 and X_1.

Step 6 : Is p-value for $X_2 < \alpha_p$? Is p-value for $X_1 < \alpha_p$?

Step 6a : If yes, include.

Step 6b : If no, exclude.

Step 7 : Is p-value for $X_1 < \alpha_p$?

Step7a : If yes, keep in model.

Step7b : If no, exclude.

Step 8 : Is p-value for $X_2 < \alpha_p$?

Step8a : If yes, keep in model.

Step8b : If no, exclude.

Finalize : End algorithm.

If the number of independent variables is large, then this procedure becomes computationally inefficient and costly. If only forward- or backward- selection is included, then this method becomes incomplete. Harrell (2015) notes that if this procedure had been proposed as a statistical technique, it would have been criticized because "it violates every principle of statistical estimation and hypothesis testing".

Recursive partitioning improves on stepwise methods by performing a stepwise method on variable spaces rather than variables. This method can be viewed as a geometrical method that splits the regions of space occupied by variables into "important" subregions where importance is calculated by some statistic. This process is done through the use of basis functions that compose the variable space. Formally, a basis of a space is a linearly independent set of vectors that span a vector space. Less formally, a basis is a set of vectors where all points in a specified space can be uniquely composed as a linear combination of these vectors. In simple OLS regression with one independent variable, x, and one dependent variable, y, the vector space of interest can be the whole xy-plane. A basis of this vector space is the x-axis and the y-axis.

Let $B_j(x)$ be a basis function. Then the goal is to approximate some function of x by $\hat{f}(\mathbf{x}) = \sum_{j=1}^{J} a_j B_j(\mathbf{x})$, where $B_j(\mathbf{x})$ can be an indicator function, a spline, or some other representation and a_j is a coefficient akin to a regression coefficient. Recursive partitioning

takes the whole variable space and finds split points t^* for each independent variable $x_i \in \mathbf{x}$ that minimize a lack-of-fit function (such as residual sums of squares). Note that if the basis functions are splines such as B-splines, the set of t^* would represent a knot sequence. Basis functions are added or subtracted for the space until a stopping rule is met. The aim of recursive partitioning is not only to find the appropriate basis functions to approximate the variable space, but also to estimate the coefficients that best fit the data. While the quality of the recursive partioning estimator depends on the specifications made (e.g., whether indicator functions or B-splines are used), these methods are typically found to have superior qualities to traditional stepwise regression methods. Friedman (1991) and Kooperberg, Stone, & Truong (1995b) prove asymptotic properties of certain recursive partitioning methods. Examples of recursive partioning in statistics occur in classification and regression trees (Breiman, Friedman, Olshen, & Stone, 1984), multivariate adaptive regression splines (Friedman, 1991), and hazard regression (Kooperberg, Stone, & Truong, 1995a).

2.8 Levenberg-Marquardt Algorithm

The weighted residual sum of squares is a common objective function to minimize in curve-fitting problems. Given a dependent variable y dependent on a set of independent variables $\mathbf{x}$, the weighted residual sum of squares is defined as

$$\begin{aligned} RSS_\sigma &= \sum_{i=1}^{k} \left[\frac{y(x_i) - \hat{y}(x_i)}{\sigma_{y_i}} \right]^2 \\ &= (\mathbf{y} - \hat{\mathbf{y}})^T \mathbf{W} (\mathbf{y} - \hat{\mathbf{y}}) \\ &= \mathbf{y}^T \mathbf{W} \mathbf{y} - 2\mathbf{y}^T \mathbf{W} \hat{\mathbf{y}} + \hat{\mathbf{y}}^T \mathbf{W} \hat{\mathbf{y}} \end{aligned} \tag{2.15}$$

where σ is the weight and typically refers to the standard error, and $\hat{y}$ is the estimated value of y. Note that if this were unweighted (i.e. $\sigma = 1$), then equation (2.15) reduces to the typical residual sum of squares. Any choice of weight could be used. If the function of $\hat{y}(x)$ is nonlinear, then the curve fitting procedure must be done iteratively.

The Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963) minimizes RSS_σ by combining two other minimization algorithms: the gradient descent method and the Gauss-Newton method. Levenberg-Marquardt adaptively switches between these two methods depending on a damping parameter λ. Prior to deriving λ, consider these two minimization methods separately.

The derivative of equation (2.15) provides the slope of the tangent line of RSS_σ. RSS_σ is minimized when this derivative is 0. Let β represent the parameter that associates $\mathbf{x}$ to $\mathbf{y}$. Then,

$$\begin{aligned}\frac{\partial}{\partial\beta}SSE_\sigma &= 2(\mathbf{y}-\hat{\mathbf{y}}(\beta))^T\mathbf{W}\frac{\partial}{\partial\beta}(\mathbf{y}-\hat{\mathbf{y}}(\beta)) \\ &= -2(\mathbf{y}-\hat{\mathbf{y}}(\beta))^T\mathbf{W}\frac{\partial\hat{\mathbf{y}}(\beta)}{\partial\beta} \\ &= -2(\mathbf{y}-\hat{\mathbf{y}}(\beta))^T\mathbf{W}\mathbf{J}\end{aligned}$$

where $\mathbf{J} = \frac{\partial\hat{\mathbf{y}}(\beta)}{\partial\beta}$ is the Jacobian, which essentially reflects how much the estimated value $\hat{\mathbf{y}}$ changes with respect to how much the parameter β changes. Setting this equation to 0 and solving for β is equivalent to solving for β under $(\mathbf{y}-\hat{\mathbf{y}}(\beta))^T\mathbf{W}\mathbf{J} = 0$. The primary issue is that the equation cannot be solved analytically. Therefore, rather than solving for β, set $\mathbf{h} = (\mathbf{y}-\hat{\mathbf{y}}(\beta))^T\mathbf{W}\mathbf{J}$ and iteratively change β until $\mathbf{h} \approx 0$. The gradient descent method determines $\mathbf{h}$ by updating the parameter in the "steepest" direction to 0 through a step parameter α : $\mathbf{h_{gd}} = \alpha\mathbf{J}^T\mathbf{W}(\mathbf{y}-\hat{\mathbf{y}}(\beta))$.

The Gauss-Newton method can be derived from a somewhat different approach. Using a first-order Taylor series expansion, where $\mathbf{h}$ represents the perturbation, then

$$\hat{\mathbf{y}}(\beta+\mathbf{h}) \approx \hat{\mathbf{y}}(\beta) + \frac{\partial\hat{\mathbf{y}}(\beta)}{\partial\beta}\mathbf{h} = \hat{\mathbf{y}}(\beta) + \mathbf{J}\mathbf{h}.$$

Using equation (2.15) again with $\hat{\mathbf{y}}(\beta) + \mathbf{J}\mathbf{h}$ substituted for $\hat{\mathbf{y}}$ gives

$$\mathbf{y}^T\mathbf{W}\mathbf{y} + \hat{\mathbf{y}}^T\mathbf{W}\hat{\mathbf{y}} - 2\mathbf{y}^T\mathbf{W}\hat{\mathbf{y}} - 2(\mathbf{y}-\hat{\mathbf{y}})^T\mathbf{W}\mathbf{J}\mathbf{h} + \mathbf{h}^T\mathbf{J}^T\mathbf{W}\mathbf{J}\mathbf{h}.$$

Now, rather than minimizing with respect to β, this function can be minimized with respect to $\mathbf{h}$:

$$\frac{\partial}{\partial \mathbf{h}} SSE_{\sigma}(\beta + \mathbf{h}) \approx -2(\mathbf{y} - \hat{\mathbf{y}}(\beta))^T \mathbf{W}\mathbf{J} + 2\mathbf{h}^T \mathbf{J}^T \mathbf{W}\mathbf{J},$$

giving

$$[\mathbf{J}^T \mathbf{W}\mathbf{J}]\mathbf{h}_{gn} = \mathbf{J}^T \mathbf{W}(\mathbf{y} - \hat{\mathbf{y}}(\beta)).$$

Note the similarities between the Gauss-Newton and gradient descent method.

The Levenberg-Marquardt method updates the damping parameter λ such that each step alternates between a Gauss-Newton solution and a gradient descent solution:

$$[\mathbf{J}^T \mathbf{W}\mathbf{J} + \lambda \mathbf{I}]\mathbf{h}_{lm} = \mathbf{J}^T \mathbf{W}(\mathbf{y} - \hat{\mathbf{y}}(\beta)). \tag{2.16}$$

In cases where λ is small, the algorithm tends towards Gauss-Newton. When λ is large, the algorithm tends towards gradient descent. The purpose of these alternations stems from gradient descent algorithm working well when the current estimates are far from their optimal values and the Gauss-Newton algorithm working well when current estimates are close to their optimal values. In practice, λ is intialized as a large value and decreases over the course of optimization with parameter estimate changes becoming more granular the closer they are to reaching optimality.

2.9 Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by patterns of various behaviors including deficits in social communication, presence of restricted and repetitive behaviors, and delays in cognitive or learning skills (American Psychiatric Association, 2013). ASD exhibits heterogeneous behavioral and etiologic profiles among those diagnosed (Betancur, 2011), and the development of ASD is nonlinear and dynamic (Girault & Piven, 2020). The causes of ASD are largely unknown as the etiology stems from multiple connected sources rather than one or two discriminant ones. ASD prevalence among school

children is currently estimated to be 1 in 54 (Shaw, et al., 2020), but this estimated proportion changes as the understanding of the disorder improves.

The difficulties that arise from ASD can be mitigated from early diagnosis and treatment (Eikeseth, et al., 2012; Eldevik, et al., 2012; Towle, et al., 2020), and it is possible to diagnose children using behavioral assessments within 24 months of age. Because early intervention and treatment of ASD have been shown to lead to these improvements, considerable effort into early identification, particularly in the pre-symptomatic period, has been made (Grazdzinski, et al., 2021). The motivation to identify ASD early coupled with the evidence of heterogeneity within the population has lead to longitudinal research studies that track infants through adolescence with neuroimaging (Hazlett, et al., 2012), genetic (Losh, et al., 2017), behavioral (Pugliese, et al., 2016), and hormonal (Geier & Geier, 2021) data being collected.

While the causes of ASD are largely unknown, evidence suggests a strong heritability component with a 9-fold increase in the odds of a full sibling or co-twin having ASD given the observed child has ASD (Sandin, et al., 2014). Further, ASD heritability ranged from 50% to 85% among large cohorts from five different countries (Bai, et al, 2019). Modern research design of ASD studies has reflected these discoveries with infant and childhood samples consisting primarily of three groups: high likelihood with ASD diagnosis, high likelihood without ASD diagnosis, and low likelihood. Those qualified as "high likelihood" come from families with a history of ASD diagnosis, particularly who have a sibling with ASD. Historically, "high likelihood" participants have been referred to "high risk" and may now be designated as having "elevated likelihood" or as having a "familial history" of ASD (Bottema-Beutel, et al., 2021). In addition to heritability, ASD has been shown to be comorbid with intellectual disabilities (Tonnsen, et al., 2016), Angelman syndrome (Veltman, et al., 2005), Down syndrome (Grzadzinski, et al., 2013), and Fragile X syndrome (Niu, et al., 2017).

There have been some key results found demonstrating an association between neuroanatomical features and ASD status (Girault & Piven, 2020). Individuals with ASD have been found to have brain overgrowth relative to their non-ASD counterparts, and this result has been found

among children (Xiao, et al., 2014), adolescents (Piven, et al., 1996), and adults (Piven, et al., 1992). This overgrowth has been found to be associated with the severity of social deficits in autistic children (Hazlett, et al., 2017). Accelerated expansion of the cortical surface area in the first year precede brain overgrowth in children with ASD (Hazlett, et al., 2017). While evidence exists demonstrating the association between amygdala volume and ASD status, the exact relationship has been conflicting (Girault & Piven, 2020) with some studies demonstrating an association between ASD and total amygdala volume (Schumann, et al., 2009), between severe cases of ASD and total amygdala volume (Sparks, et al., 2002), between ASD and right amygdala volume (Munson, et al., 2006), and between cases of ASD with greater joint attention and total amygdala volume (Mosconi, et al., 2009). Greater volumes of extra-axial cerebrospinal fluid have been found to be associated with ASD diagnosis and severity as early as 6-months (Shen, et al., 2013; Shen, et al., 2017).

2.10 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is a commonly used noninvasive imaging modality that produces three-dimensional anatomical images based on the nuclear magnetic resonance (NMR) phenomenon (Lauterbur, 1973). The NMR phenomenon cannot be explained exactly using classical physics (Suetens, 2009), so the following explanation is approximate to the technical details for MRI techniques. MRI measures a magnetic property of tissue. In an external magnetic field, there will be an assortment of atoms with each atom having its own angular momentum and magnetic moment. Further, each subatomic particle within each atom (i.e., the electron, proton, and neutron) has its own angular momentum. These angular momenta have associated spins, roughly equivalent to how planets spin on their axes in space. NMR evaluates the behavior of atomic nuclei with spin angular moments and associated magnetic moments in an external magnetic field.

The atomic nucleus of various isotopes have unique pairs of spin values and magnetic moments. For example, hydrogen has a spin of $\frac{1}{2}$ with associated magnetic moment of 42.6.,

nitrogen has a spin of 1 and magnetic moment of 3.1, and calcium has a spin of $\frac{7}{2}$ and magnetic moment of -2.9. Note that carbon-12 and oxygen-16, isotopes common in organic matter, have 0 spin. Half spins indicate that the nuclei have two possible states, the "spin up" state and the "spin down" state. The sum of these individual spins is called the *net magnetization* of the external magnetic field. In a typical, undisturbed field, the net magnetization of these spins is 0 because the spins of each atomic nucleus is randomly distributed in any given direction. However, when an external magnetic field is produced, these nuclei align either with the field (parallel alignment) or against the field (anti-parallel alignment).

The magnetic vector of spinning nuclei comprises two orthogonal spin components. One component is the longitudinal component (z-axis) and the other component is the transverse component (xy-plane). The rotation of the transverse component in the presence of a large external magnetic field is called precession, and its spin resembles a wobbling motion. In the presence of an external magnetic field, more spins align parallel to the field than anti-parallel. This alignment derives from the low energy state exhibited during parallel alignment. As such, the net magnetization is positive along the longitudinal component.

Radiofrequency (RF) pulses of the same frequency, or resonance, as these spins will excite the atomic nuclei from mostly parallel to mostly anti-parallel alignment (a high energy state), which rotates the longitudinal component to its antipode (i.e. spin inversion). After the excitation period, the atomic nuclei return to parallel alignments and reach equilibrium. The recovery rate of the longitudinal component of the spin is dependent on tissue characteristics. As the MRI rotates its external magnetic field around the tissue, a quadrature detector determines the time and strength of the inversion process of each *voxel*, which is a pixel with volume. The signal from the inversion recovery process produces the three-dimensional images within an MRI.

MRI of brain tissue comes primarily in two forms: structural and functional MRI. Structural MRI examines the anatomical characteristics of the brain whereas functional MRI examines brain activity. The principal measure determined by structural MRI is volume, but from volume other anatomical features can be determined such as surface area or curvature. Structural

MRI has many clinical uses such as the identification of tumors (Kucharczyk, et al., 1985) or the diagnosis of Alzheimer's disease (Wattjes, 2011). Functional MRI takes advantage of the relationship between blood oxygenation and the inversion relaxation signal to determine spatial regions of brain activity. This relationship is denoted as BOLD (blood oxygenation-level dependent) effects. When neurons activate, blood flow needs to increase to the area of activation to accommodate the need for oxygen. Therefore, changes in BOLD can be used to measure changes in neural activity (Logothetis, Guggenberger, Peled, & Pauls, 1999). In a clinical setting, fMRI can be used to assess strokes or clots (Crofts, Kelly, & Gibson, 2020), but much of the value of fMRI has come from its impact on cognitive neuroscience (Matthews, Honey, & Bullmore, 2006).

CHAPTER 3: THE JOINT LATENT CLASS MIXED MODEL WITH HAZARD REGRESSION

3.1 Hazard Regression

In survival analysis, the proportional hazards model is a commonly used semiparametric method whose advantages include the linear interpretability of effects and the minimal assumptions needed to model the hazard function. Proportional hazards models have the ability to estimate the effects of the covariates irrespective of knowledge of the baseline hazard. While the baseline hazard does not need to be known in a proportional hazards model, proportional hazards models do require the proportionality assumption. Hazard regression (HARE) uses an adaptive regression technique to estimate the conditional log hazard function from right-censored survival data with multiple covariates (Kooperberg, Stone, & Truong, 1995). The conditional log-hazard function is estimated by an adaptive method similar to Friedman's (1991) multivariate adaptive regression spline method. Given a p-dimensional linear space of positive functions and a set of basis function $B_1, \ldots, B_p$, HARE estimates the conditional log-hazard function $\alpha(t|\mathbf{x})$ by maximizing the partial likelihood under $\beta = (\beta_1, \ldots, \beta_p)$ such that

$$\alpha(t|\mathbf{x}, \beta) = \sum_{j=1}^{p} \beta_j B_j(t|\mathbf{x}), t \geq 0. \tag{3.17}$$

The allowable basis functions are B-spline bases in the covariates, B-spline bases in time t, and their tensor products $B_i \otimes B_j$. Both the basis space and its dimension, p, are determined adaptively using recursive partitioning. HARE has two advantages for modeling the log-hazard function: (i) the estimation is entirely data-driven with no parametric assumptions about the distribution of $\alpha(t|\mathbf{x})$ and (ii) the proportionality assumption is relaxed. The latter advantage is

implemented by allowing the log-hazard function to include interactions between time and the covariates.

Let $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. T_i is the time to failure for individual i, C_i is the time to censoring for individual i, and δ_i is the censoring indicator for individual i. The censoring indicator $\delta_i = 1$ if failure is observed and $\delta_i = 0$ if censoring is observed. The partial likelihood (Miller, 1981) for the observed time to event y_i given covariates $\mathbf{x}$ and under parameters β is $\prod_{i=1}^{n}[f(y_i|\mathbf{x}_i,\beta)]^{\delta_i}[1-F(y_i|\mathbf{x}_i,\beta)]^{1-\delta_i}$. This implies the log-likelihood is

$$\begin{aligned}
\phi(y_i,\delta_i|\mathbf{x}_i,\beta) &= \log\left[\prod_{i=1}^{n}[f(y_i|\mathbf{x}_i,\beta)]^{\delta_i}[1-F(y_i|\mathbf{x}_i,\beta)]^{1-\delta_i}\right] \\
&= \sum_{i=1}^{n}\log\left[[f(y_i|\mathbf{x}_i,\beta)]^{\delta_i}\right] + \log\left[[1-F(y_i|\mathbf{x}_i,\beta)]^{1-\delta_i}\right] \\
&= \sum_{i=1}^{n}\delta_i \log f(y_i|\mathbf{x}_i,\beta) + (1-\delta_i)\log(1-F(y_i|\mathbf{x}_i,\beta)) \\
&= \sum_{i=1}^{n}\delta_i \log f(y_i|\mathbf{x}_i,\beta) - \delta_i \log(1-F(y_i|\mathbf{x}_i,\beta)) + \log(1-F(y_i|\mathbf{x}_i,\beta)) \\
&= \sum_{i=1}^{n}\delta_i \left[\log f(y_i|\mathbf{x}_i,\beta) - \log(1-F(y_i|\mathbf{x}_i,\beta))\right] + \log(1-F(y_i|\mathbf{x}_i,\beta)) \\
&= \sum_{i=1}^{n}\delta_i \left[\log \frac{f(y_i|\mathbf{x}_i,\beta)}{(1-F(y_i|\mathbf{x}_i,\beta))}\right] + \log(1-F(y_i|\mathbf{x}_i,\beta)) \\
&= \sum_{i=1}^{n}\delta_i \log\lambda(y_i|\mathbf{x}_i,\beta) + \log(1-F(y_i|\mathbf{x}_i,\beta)) \\
&= \sum_{i=1}^{n}\delta_i \alpha(y_i|\mathbf{x}_i,\beta) + \log\exp(-\int_0^{y_i}(\exp(\alpha(u_i|\mathbf{x}_i,\beta))))du \\
&= \sum_{i=1}^{n}\delta_i \alpha(y_i|\mathbf{x}_i,\beta) - \int_0^{y_i}\exp(\alpha(u_i|\mathbf{x}_i,\beta))du, y \geq 0, \delta_i \in \{0,1\} \qquad (3.18)
\end{aligned}$$

where the derivation of equation (3.2) is determined by the relationships between survival functions

$$e^{\alpha(t|\mathbf{x})} = \lambda(t|\mathbf{x}) = \frac{f(t|\mathbf{x})}{1-F(t|\mathbf{x})}, \quad 1-F(t|\mathbf{x}) = e^{-\int_0^t \lambda(u|\mathbf{x})}du = e^{-\int_0^t e^{\alpha(u|\mathbf{x})}}du,$$

$$f(t|\mathbf{x}) = e^{\alpha(u|\mathbf{x})} e^{-\int_0^t e^{\alpha(u|\mathbf{x})}} du.$$

Maximizing this partial log-likelihood resolves to solving for β in equation (3.1). The corresponding score function $\mathbf{S}(\beta)$ and Hessian matrix $\mathbf{H}(\beta)$ are respectively determined by the first derivative of the partial log-likelihood function with respect to β_j and the $p \times p$ matrix of second derivatives of the log-likelihood function with respect to β_j, β_k.

$$\begin{aligned}
\mathbf{S}(\beta) &= \frac{\partial}{\partial \beta_j} \phi(y, \delta|\mathbf{x}, \beta) \\
&= \delta B_j(y|\mathbf{x}) - \int_0^y B_j(u|\mathbf{x}) \exp(\alpha(u|\mathbf{x}, \beta)) du, \quad 1 \leq j \leq p, \quad y \geq 0, \quad \delta \in \{0, 1\} \\
\mathbf{H}(\beta) &= \frac{\partial^2}{\partial \beta_j \partial \beta_k} \phi(y, \delta|\mathbf{x}, \beta) \\
&= -\int_0^y B_j(u|\mathbf{x}) B_k(u|\mathbf{x}) \exp(\alpha(u|\mathbf{x}, \beta)) du, \quad 1 \leq j, \quad k \leq p, \quad y \geq 0, \quad \delta \in \{0, 1\}.
\end{aligned}$$

The Newton-Raphson method is used to estimate $\hat{\beta}$ with an initial guess for $\hat{\beta}^{(0)}$ and $\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} - 2^{-\nu}[\mathbf{H}(\hat{\beta}^{(m)})]^{-1}\mathbf{S}(\hat{\beta}^{(m)})$ where ν is a step-halving constant. The iterations stop when the difference between repeated log-likelihood calculations is $< 10^{-6}$. The goal, then, is to determine the allowable basis functions with attention to knot placement to estimate $\alpha(t|\mathbf{x})$.

The allowable spaces of basis functions for HARE, $\mathcal{G}$, are linear in order to minimize numerical integrations over the knot sequence of $\mathbf{t} := (t_1, t_2, \ldots, t_k)$ (Kooperberg, Stone, & Truong, 1995a) and because an approximated function composed of appropriately chosen allowable linear spaces converges in mean to a generalized linear model (Stone, 1994; Kooperberg, Stone, & Truong, 1995b). Formally, the allowable spaces $G \in \mathcal{G}$ in HARE are defined as follows:

- There is only one $G \in \mathcal{G}$ with minimal dimension p_{min},
- Each $G \in \mathcal{G}$ is a linear space having dimension $p \geq p_{min}$,
- If $G \in \mathcal{G}$ has dimension $p > p_{min}$, then there is at least one subspace $G_0 \in \mathcal{G}$ of G with dimension $p - 1$,

- If $G_0 \in \mathcal{G}$ has dimension p, then there is at least one space $G \in \mathcal{G}$ with dimension $p+1$ whose subspace is G_0.

Less formally, this means that HARE determines B_j and estimates its corresponding β_j in a stepwise fashion where tensor products of spaces require each individual space. Let k represent a knot along a knot sequence. Then HARE has basis functions of the form $1, (t_k - t)_+, x_m, (x_{mk} - x_m)_+, x_m x_n, (t_k - t)_+ x_m, (t_k - t)_+ (x_{mk} - x_m)_+, x_m (x_{nk} - x_n)_+, (x_{mk} - x_m)_+ x_n$, and $(x_{mk} - x_m)_+ (x_{nk} - x_n)_+$, where x_m, x_n are separate covariates, t_k is a knot in time, $x_{\cdot k}$ is a knot in a covariate, and $(\cdot)_+$ represents the positive part of the function. The tensor products $x_m x_n, (t_k - t)_+ x_m, (t_k - t)_+ (x_{mk} - x_m)_+, x_m (x_{nk} - x_n)_+, (x_{mk} - x_m)_+ x_n$, and $(x_{mk} - x_m)_+ (x_{nk} - x_n)_+$ are allowable only if the individual basis functions are included. For example, if x_1, x_2, and $(7-t)_+$ are included in the estimate of $\alpha(t|\mathbf{x})$, but x_3 is not, then $x_1 x_2, x_1 (7-t)_+,$ and $x_2 (7-t)_+$ are allowable, but $x_3 (7-t)_+$ is not.

HARE begins the partitioning procedure of determining the allowable spaces from $\mathcal{G}$ with the minimal space $G_{min} = 1$ (the constant space). That is, HARE always initializes its estimation of $\alpha(t|\mathbf{x}) = \beta_1(1) = \beta_1$. HARE then proceeds with adding new spaces $G \in \mathcal{G}$ where each $(p-1)$-dimensional space G_0 is replaced by a p-dimensional space G that includes G_0 as a subspace. When determining the new G allowable space, candidate basis functions B_j include linear covariates, a new knot in time, a new knot in a covariate, and a tensor product of two existing basis functions from G_0. Note from the list of basis function forms above that $B_j = t$ is not allowed. This exclusion stems from the need of a constant tail for the cumulative hazard (i.e. $\Lambda(t|\mathbf{x}) = \infty$ when $t \to \infty$) and conversely the need for a constant tail for the survival function (i.e. $S(t|\mathbf{x}) = 0$ when $t \to \infty$). Details on this constraint can be found in Kooperberg, Stone, & Truong (1995a) and Stone, et al. (1997). Example 1 provides an illustration of the addition of basis functions for G.

G is determined from finding the candidate basis function that maximizes the Rao statistic. Let $\hat{\beta}^{(0)}$ be the maximum likelihood estimate of β corresponding to space G, and let β_p be defined as the coefficient for the basis function needed to go from subspace G_0 to space G.

The Rao statistic is a score statistic calculated by $R = \mathbf{S}(\hat{\beta}_p^{(0)})/\sqrt{\mathbf{I}^{-1}(\hat{\beta}^{(0)})_{pp}}$, where $\mathbf{S}(\hat{\beta}_p^{(0)})$ is the score function evaluated at the maximum likelihood estimate in p-dimensional space, and $\mathbf{I}^{-1}(\hat{\beta}^{(0)})$ is the observed Fisher information matrix observed at the maximum likelihood estimate of β. The addition of new G allowable spaces follows this algorithm:

- Calculate Rao statistic for all spaces obtained from G_0 by adding a basis function $B_{l0}(x_l) = x_l$ to G_0
- Calculate Rao for all allowable spaces obtained from G_0 by adding a basis function to G_0 comprising a tensor product of two tensor functions in G_0
- Calculate Rao statistic for a space obtained from G_0 by adding a basis function constructed by adding a new knot in t
- Calculate Rao statistic for a space obtained from G_0 by adding a basis function constructed by adding a new knot in covariate m
- Select space G that maximizes the absolute value of the Rao statistic.

After each space G is determined, the BIC (Schwarz, 1978) is stored for model selection procedures. Candidate basis functions are no longer added when either (a) the number of basis functions included is $\min(6n^{1/5}, n/4, 50)$, (b) the change in the maximized log-likelihood function is $< \frac{1}{2}(P-p) - \frac{1}{2}$ where P is the number of basis functions and p is the dimensionality of G, or (c) the algorithm yields no possible new basis function (Kooperberg, Stone, & Truong, 1995a).

Example 1. Consider a log hazard function with candidate basis functions x_1 and t, where x_1 is a continuous covariate. The addition phase of the estimation of $\alpha(t|\mathbf{x})$ would begin with $G_{min} = 1$. The following G could add either x_1 or $(t_{k1} - t)_+$ for knot t_{k1}. After determining G that maximizes the Rao statistic, this G is set to G_0 and the following G could add x_1, $(t_{k2} - t)_+$, $(x_1 - x_{1k1})_+$, or $x_1(t_{k1} - t)_+$ depending on what basis functions exist in G_0. □

Knot selection for candidate basis functions requires special attention. The knot sequence for t or for a covariate x is determined adaptively through the minimization of the BIC, which is defined by $BIC = n \log \frac{RSS}{n} + p \log n$ where RSS is the residual sum of squares $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and p is the number of parameters in the model. Knots are added sequentially within one basis function until a minimal RSS is determined. Figure 3.1.1 illustrates the process of knot addition for an arbitrary length of time. Note that the BIC is saved for each knot addition. After the addition phase has concluded, knots are deleted from the basis function until the minimum model size has been reached. BIC is calculated at each step during the deletion phase as well. The optimal basis function with knot sequence $\mathbf{t}_k$ is the one with the minimum BIC value. BIC is used in lieu of RSS since RSS increases arbitrarily with each additional variable. Figure 3.1.2 illustrates the BIC selection procedure for adaptive knot selection.

Following the addition phase of $G \in \mathcal{G}$ is the deletion phase, which carries out the candidate basis function algorithm above with two features changed: (i) the deletion phase goes from space G to space G_0, and (ii) Wald statistic $\hat{\beta}_p / SE(\hat{\beta}_p)$ is used instead of the Rao statistic. This latter modification derives from the fact that the Rao statistic is based on the maximum likelihood estimate in G_0 whereas Wald is based on G space (Kooperberg, Stone, & Truong; 1995a). After each model has been estimated through the iterative addition and deletion phases of G construction, the model that minimizes BIC is the final estimate of the conditional log-hazard function $\alpha(t|\mathbf{x}, \beta)$.

3.2 The Joint Latent Class Mixed Model

Maximum Likelihood Estimation

The joint latent class mixed model (JLCMM) is composed of three component models: a latent class membership model estimated from multinomial logistic regression, a longitudinal trajectory model estimated from a linear mixed-effects model, and a time-to-event model estimated by a hazard function. Let the sample comprise $i = 1, \ldots, N$ total participants. The JLCMM is estimated by maximizing the log-likelihood $\ell(\theta_K)$ of the data for a fixed number

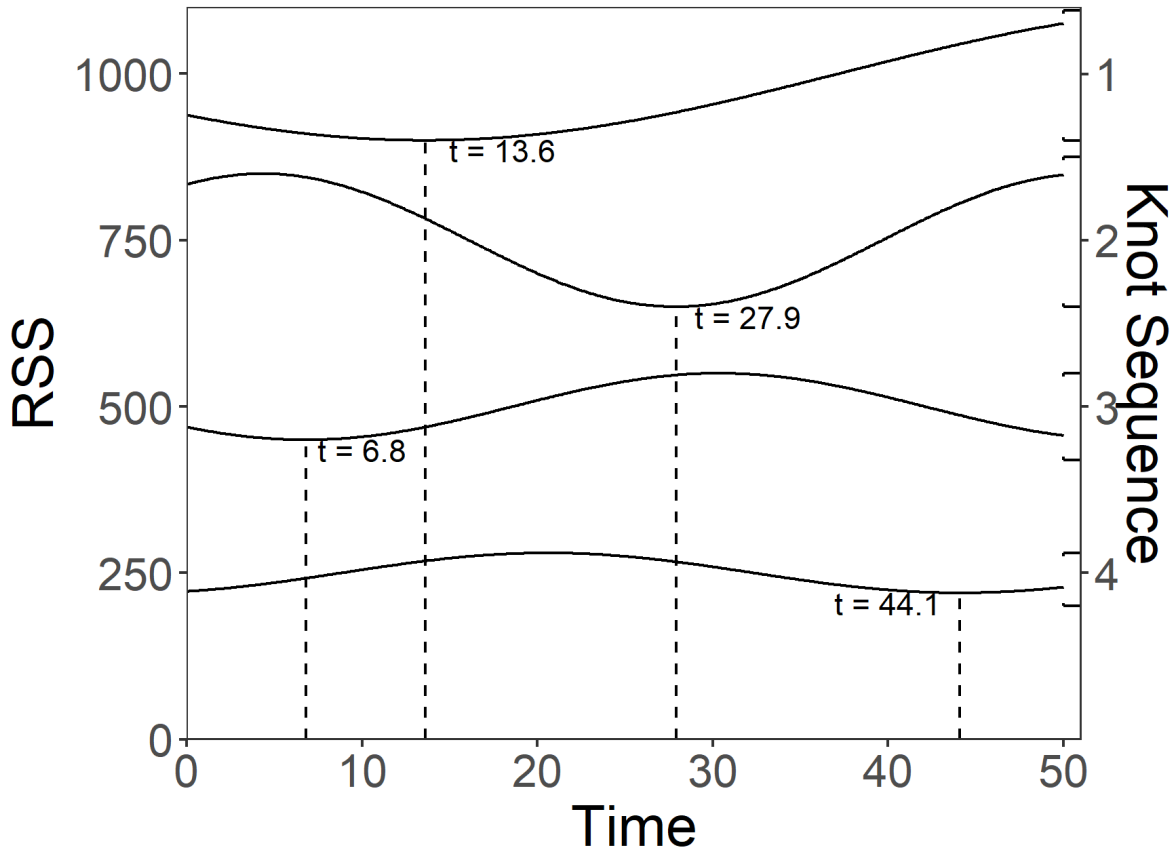

Figure 3.1.1 RSS Calculation During Knot Addition Phase. Dashed lines show minimized RSS based on each knot location t_k.

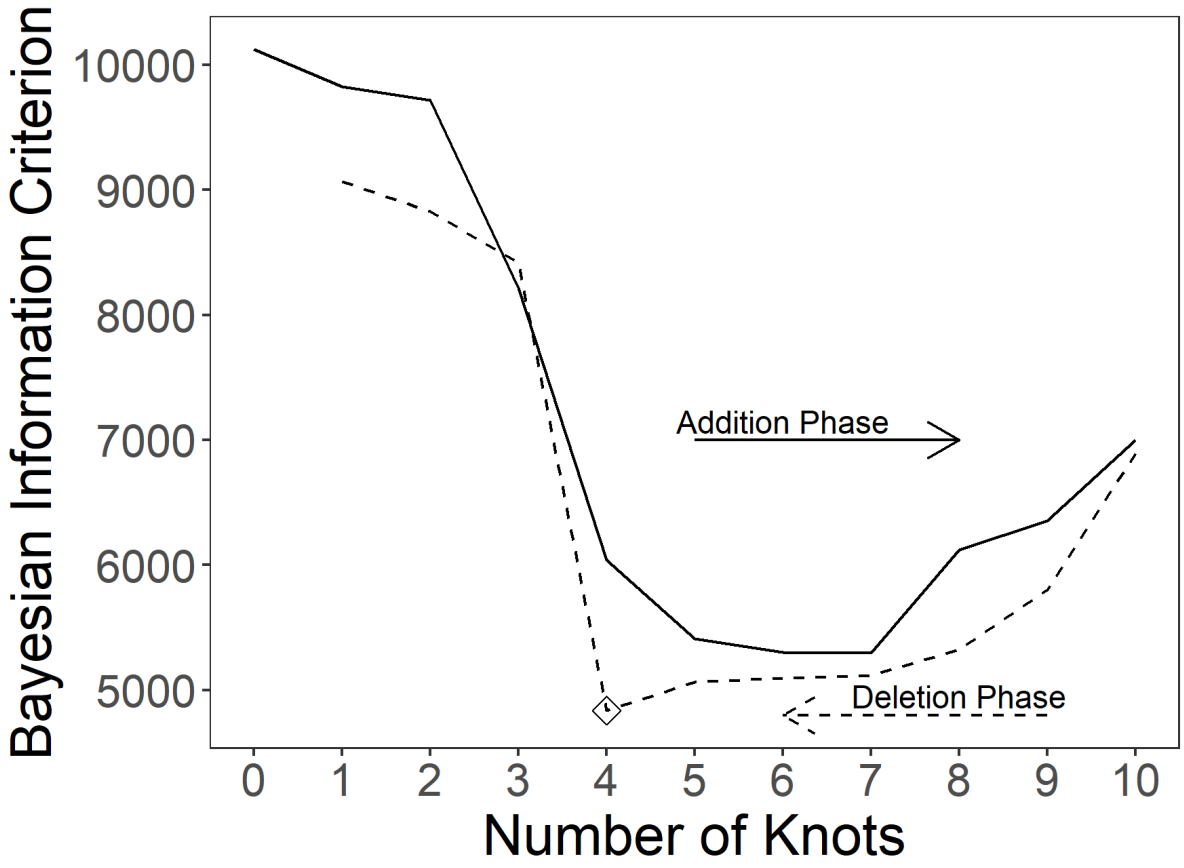

Figure 3.1.2 Knot Determination from BIC Calculations during Addition and Deletion Phases of Adaptive Knot Placement. The optimal model is shown with a diamond, where the total number of knots within this space is 4.

of classes $k = 1, \ldots, K$ over the entire set of parameters for the K latent classes, θ_K. This log-likelihood is defined in equation (3.3):

$$\ell(\theta_K) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_{ik} f(Y_i|c_i = k, \theta_K) \lambda_i(T_i|c_i = k, \theta_K)^{\delta_i} S_i(T_i|c_i = k, \theta_K). \tag{3.19}$$

π_{ik} represents the latent class membership probability, $f(\cdot)$ represents the longitudinal trajectory density function, and $\lambda_i(\cdot)$ and $S(\cdot)$ represent the time-to-event hazard and survival functions, respectively. Note that these latter two functions correspond to the determination of $\phi_i(\cdot)$ in equation (3.2). The density $f(\cdot) \sim N(\mathbf{X}\beta_k + \mathbf{Z}b_k, \mathbf{Z}\mathbf{G_k}\mathbf{Z}^T + \mathbf{R})$ for each latent class k. Each model has sample-wide and class-specific estimates for each latent class k.

The log-likelihood $\ell(\theta_K)$ is estimated iteratively using the Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963) with strict convergence under three criteria: parameter stability ($\sum_{j=1}^{p}(\theta_K(j)^{(l)} - \theta_K(j)^{(l-1)})^2 \leq \epsilon_a$), $\ell(\theta_K)$ stability ($\ell(\theta)^{(l)} - \ell(\theta)^{(l-1)} \leq \epsilon_b$), and partial derivative size ($\frac{\nabla(\ell(\theta_K^{(l)}))^T \mathcal{H}^{(l)-1} \nabla(\ell(\theta_K^{(l)}))^T}{p} \leq \epsilon_c$) where each threshold $\epsilon = 10^{-4}$. (Proust-Lima, et al., 2017). This estimation procedure has been validated for the purposes of identifiability in finite mixture models by Proust-Lima et al (2009) using simulation studies. After assessing a predetermined number of latent classes, such as 1 - 5, the optimal number of latent classes is selected as the model which minimizes the BIC (Schwarz, 1978).

Latent Class Membership Probability

Let the sample of $i = 1, \ldots, N$ participants be divided into $k = 1, \ldots, K$ homogeneous latent subgroups called latent classes. Latent class membership for participant i, c_i, is determined from a mulitnomial logistic regression model conditioned on a set of $X_{(c)}$ that may be unique to the latent class membership model or common to other components of the JLCMM. Let π_{ik} denote the probability that participant i belongs to latent class k. Then, this model is estimated by

$$\pi_{ik} = P(c_i = k|X_{(c)}) = \frac{e^{\xi_{0k} + X_{(c)}^T \xi_{1k}}}{\sum_{l=1}^{K} e^{\xi_{0l} + X_{(c)}^T \xi_{1l}}} \tag{3.20}$$

where ξ represents the coefficients associated with latent class k. For identifiability, the largest latent class is the referent class (i.e. $\xi_{0K} = 0, \xi_{1K} = 0$). An assumption of the JLCMM is that given knowledge about latent class membership, the longitudinal outcome and time-to-event are independent. That is, a participant's longitudinal trajectory and event hazard are class-specific where knowledge of class membership dictates these two outcomes.

Class-Specific Longitudinal Outcome

Conditional on the latent class k, the longitudinal outcome described at measurement times t_{ij} for participant i and measurement j is estimated using a linear mixed-effects model

$$Y_i(t_{ij})|_{c_i=k} = \mathbf{X}\beta_k + \mathbf{Z}b_{ik} + \epsilon_i(t_{ij}) \tag{3.21}$$

where $\mathbf{X} = X_{li}(t_{ij})$ is a vector of potentially time-dependent fixed effects covariates associated with the coefficients for latent class k, β_k. Similarly, $\mathbf{Z} = Z_i(t_{ij})$ is a vector of potentially time-dependent random effects covariates associated with the coefficients for latent class k, b_k. $\mathbf{G}$ and $\mathbf{R}$ can follow any typical covariance structures, but typically $\mathbf{G}$ is unstructured and $\mathbf{R} = \sigma^2 I_{n_i}$ $(j = 1, \ldots, n_i)$ in order to reduce the number of estimated parameters and for helping identifiability. $\epsilon_i(t_{ij})$ can represent independent errors at each time t_{ij} or represent a correlation process such as autoregressive correlation. Covariates $\mathbf{X}$ and $\mathbf{Z}$ may overlap but do not necessarily have to.

Class-Specific Hazard Function using HARE

From equation (3.2),

$$\begin{aligned}\phi_i(T_i, \delta_i) &= \log \lambda_i(T_i, \delta_i) S_i(T_i, \delta_i) \\ &= \delta_i \alpha(\tilde{T}_i | X_{(s)}, \beta) - \int_0^{\tilde{T}_i} \exp(\alpha(u_i | X_{(s)}, \beta)) du, y \geq 0, \delta_i \in \{0, 1\}\end{aligned}$$

where $\tilde{T}_i$ is the observed time, $X_{(s)}$ are the covariates associated with the time-to-event model, and $\alpha(t|X_{(s)}) = \sum_{j=1}^{p} \beta_j B_j(t|X_{(s)})$ is the log-hazard function as estimated by HARE. Con-

ditioned on the latent class k, the proportional hazards model estimated within a JLCMM is

$$\lambda_i(t|c_i = k, \tilde{\theta}_k) = \lambda_{0k}(t)e^{X_{(s)}^T\tilde{\theta}_k}. \tag{3.22}$$

where $\tilde{\theta}_k$ are the class-specific parameters associated with $X_{(s)}$. With HARE, equation (3.6) becomes

$$\begin{aligned}\lambda_i(t|c_i = k, \tilde{\theta}_k) &= \exp\{\alpha(t|c_i = k, \tilde{\theta}_k)\} \\ &= \exp\{\sum_{k=1}^{K}\sum_{j=1}^{p}\beta_{jk}B_j\}.\end{aligned}$$

Conditional Independence

A central assumption for the JLCMM is that all of the information about the longitudinal trajectory and time-to-event outcome are contained within a latent class, which indicates that these two outcomes are conditionally independent once latent class assignment is determined from the model. This assumption allows the terms in equation (3.3) to be separable. Therefore, it is important to establish this assumption when estimating JLCMMs in order to justify the latent class assignments. Jacqmin-Gadda et al. (2010) offered a score test to test whether residual independence between the two outcomes was present after latent class assignment was determined. The alternative hypothesis for this score test $\mathcal{H}_1$ assumes a random effect in equation (3.6)

$$\lambda_i(t|c_i = k, X_{(s)}, \tilde{\theta}_k) = \lambda_{0k}(t)e^{X_{(s)}^T\tilde{\theta}_k + b_{ik}^T\eta}$$

whose parameter η is a $p-$dimensional vector of random effects from the longitudinal model for latent class k, b_{ik}. The hypothesis test assesses $\mathcal{H}_0 : \eta = 0$ vs. $\mathcal{H}_1 : \eta \neq 0$ using a score test $U = \sum_{i=1}^{N}\sum_{k=1}^{K}\hat{\pi}_{ik}(E_i - \Lambda_{ik}(T_i))\hat{b}_{ik}$, where E_i is an indicator function determining whether the observed event time is less than censoring time for individual i (i.e., $T_i \leq C_i$) and Λ_{ik} is the $k-$specific cumulative hazard function. Under $\mathcal{H}_0$, $U^T Var(U)^{-1}U \sim \chi_p^2$ where rejection

of $\mathcal{H}_0$ implies a residual relationship between the longitudinal outcome and time-to-event exist within a latent class. This $\eta \neq 0$ conclusion indicates that conditional independence has been violated making the JLCMM suspect of inference. This score test has been shown to be a more powerful method of detecting departure from conditional independence than other existing methods (Jacqmin-Gadda, et al., 2010; Proust-Lima, et al., 2014).

Posterior Classification of Latent Class Membership

One of the motivations for using JLCMMs, or any latent class model in general, is the ability to estimate probabilities of latent class membership within the sample. JLCMMs estimate this probability using Bayes' Theorem to calculate the posterior probability of latent class membership conditioned on the joint outcomes of the model. Letting Y_i be the longitudinal outcome, $(\tilde{T}_i, \delta_i)$ be the time-to-event outcome, and $\hat{\theta}_K$ be the estimates of all the parameters, then the posterior probability of latent class membership for individual i in class k is

$$\begin{aligned}\hat{\pi}_{ik}^{(Y,\tilde{T})} &= P(c_i = k | Y_i, (\tilde{T}_i, \delta_i), \hat{\theta}_k) \\ &= \frac{\hat{\pi}_{ik} f(Y_i | c_i = k, \hat{\theta}_K) \lambda_i(\tilde{T}_i | c_i = k, \hat{\theta}_K)^{\delta_i} S_i(\tilde{T}_i | c_i = k, \hat{\theta}_K)}{\sum_{l=1}^{K} \hat{\pi}_{il} f(Y_i | c_i = l, \hat{\theta}_K) \lambda_i(\tilde{T}_i | c_i = l, \hat{\theta}_K)^{\delta_i} S_i(\tilde{T}_i | c_i = l, \hat{\theta}_K)}. \end{aligned} \tag{3.23}$$

Note that these posterior probabilities are estimated after estimates for $\pi_{ik}, f(Y_i), \lambda_i(\tilde{T}_i), S_i(\tilde{T}_i)$ are determined.

Latent class membership is determined modally, where $\hat{c}_i^{(Y,\tilde{T})} = \underset{k}{\operatorname{argmax}} (\hat{\pi}_{ik}^{(Y,\tilde{T})})$. Modal assignment provides utility in inference, but it is not technically necessary, and some situations may exist when dimensional assignment to classes may prove more useful (Vermunt, 2010). If modal assignment is the intention, then a more discriminatory posterior classification indicates a better fitting model. This discrimination can be determined in a few ways such as through using proportional thresholds or through a posterior classification table.

Proportional thresholds gauge discrimination by determining the proportion of individuals assigned to a latent class who have a posterior probability greater than some threshold value. For example, if the threshold is 0.7, then a proportion $P(\hat{c}_{i1}^{(Y,\tilde{T})} \geq 0.7) = 0.6563$ indicates that

almost 66% of the sample assigned to latent class 1 had a posterior probability of assignment to LC 1 greater than 0.7. Greater proportions indicate greater discrimination between the latent class assignments. Higher thresholds indicate stricter guidelines for determining discrimination. Thresholds typically used include 0.7, 0.8, and 0.9 (Proust-Lima, et al., 2014).

Similarly, a posterior classification table of assignment in each latent class may be used to assess discrimination. This $K \times K$ table provides the mean probability of assignment in each latent class (along the columns) given modal assignment (along the rows). For example, cell (1,1) would indicate the mean posterior probability of assignment into LC 1 given modal assignment to LC 1 and cell (2,3) would indicate the mean posterior probability of assignment into LC 3 given modal assignment to LC 2. A highly discriminatory assignment would have its diagonal terms close to 1 and its off-diagonal terms close to 0.

Fitted Values

The linear-mixed effects model within the JLCMM provides subject-specific, sample-specific, and class-specific predicted values. The former two predictions can be compared with observed data to determine model fitness. The latter can be used for model inference, but obviously not for comparison with some observed classes. Consider participant i with time j assigned to class k. Then the marginal, class-specific prediction for Y is

$$\hat{Y}_{ijk} = Z_i(t_{ij})^T \hat{b}_k + \mathbf{X}(t_{ij})^T \hat{\beta}_k$$

and the subject-specific prediction for individual i in LC k is

$$\hat{Y}_{ijk} = Z_i(t_{ij})^T (\hat{b}_k + \hat{G}_k Z_i^T \hat{V}_{ik}^{-1}(Y_i - \mathbf{X}\hat{\beta}_k - Z_i \hat{b}_k)) + \mathbf{X}(t_{ij})^T \hat{\beta}_k.$$

Individual-level or class-level predictions can be averaged by taking the weighted mean of these outcomes with weights being equal to probabilities of latent class assignments. The subject-specific and sample-specific residuals can be computed by subtracting the predicted

values ($\hat{Y}_{ij}$) from the observed values (Y_{ij}) to help assess model fit. Model fitness can also be assessed from the survival functions $\hat{S}_k(t) = \sum_{i=1}^{N} \hat{\pi}_{ik} S_i(t|c_i = k, \hat{\theta}_K)$.

Individual Dynamic Prediction

With joint models having been developed under a clinical framework, much research has been dedicated to developing predictive tools that allow for dynamic prognostic decision making (Proust-Lima & Taylor, 2009; Taylor, Yu, & Sandler, 2005; Rizopolous, 2011). One such tool for individual-level dynamic prediction incorporates covariates and longitudinal outcomes measured until time s to predict the probability of an event occurring within a time window $[s, s+t]$. These non-negative parameters are often denoted as the time at prediction (s) and the horizon (t) (Proust-Lima, et al., 2014). While t is chosen by the investigator, it should reflect a reasonable margin in time that reflects the rate of event occurrence. An horizon too large where many events occur or too small where very few occur may prove useless for prediction. Let $Y_i^{(\leftarrow S)}$ denote the longitudinal outcome until the time at prediction. $\mathbf{X}$ is the design matrix of covariates, T_i is the time-to-event, and θ is the vector of parameters for the JLCMM. Then these predicted probabilities are (Proust-Lima, et al., 2014)

$$
\begin{aligned}
P(T_i \leq s+t | T_i \geq s, Y_i^{(\leftarrow S)}, \mathbf{X}, \theta) = & \\
& \sum_{k=1}^{K} P(T_i \leq s+t | T_i \geq s, c_i = k, \mathbf{X}, \theta) \times P(c_i = k | T_i \geq s, Y_i^{(\leftarrow S)}, \mathbf{X}, \theta) \\
& = \frac{\sum_{k=1}^{K} \pi_{ik} f(Y_i^{(\leftarrow S)} | c_i = k, \mathbf{X}\theta)(S_i(s|c_i = k, \mathbf{X}\theta) - S_i(s+t|c_i = k, \mathbf{X}\theta))}{\sum_{k=1}^{K} f(Y_i^{(\leftarrow S)} | c_i = k, \mathbf{X}\theta) S_i(s|c_i = k, \mathbf{X}\theta)}
\end{aligned}
\tag{3.24}
$$

As expected, these predictions can be made using estimates $\hat{\theta}$ in place of θ. The variance of these predictions are not analytic (Proust-Lima & Taylor, 2009; Rizopolous, 2011), so confidence bands need to be estimated numerically using a method such as Monte Carlo or bootstrapping. The draws for $\hat{\theta}$ are assumed to be asymptotically normal (Proust-Lima, et al., 2014).

Predictive Accuracy using Expected Prognostic Observed Cross-Entropy

Given a conditional density $f_{\tilde{T}|Y}$ where $\tilde{T}$ is the observed time-to-event, the expected prognostic observed cross-entropy (EPOCE) is defined as $E(-\log f_{\tilde{T}|Y,Y^{(\leftarrow S)}}|T \geq s)$ (Commenges, Liquet, Proust-Lima; 2012). The EPOCE is an information theory-based predictive fit estimate that estimates the expected risk (in a loss function sense) of assuming the form of $f_{\tilde{T}|Y}$ to predict time-to-event as an estimate for $f_{\tilde{T}|Y}$, similar to other cross-entropy functions. EPOCE can be calculated using leave-one-out cross validation using equation (3.9):

$$\text{CVPOL}_a(s) = -\frac{1}{N_s}\sum_{i=1}^{N_s} F_i(\hat{\theta}, s) + N \text{ Trace}(H^{-1}Q_s), \tag{3.25}$$

where N_s are the number of participants still at risk at time s, H is the Hessian matrix of equation (3.3), and $Q_s = \frac{1}{N_s(N-1)}\sum_{i=1}^{N} I(\tilde{T}_i \geq s)\hat{v}_i(s)\hat{d}_i^T$ for some gradients of the conditional log-likelihood, v and d. F_i is determined as the log-probability of equation (3.8). EPOCE has been shown in simulations (Proust-Lima, et al., 2014; Commenges, Liquet, Proust-Lima, 2012) to be advantageous over other methods such as quadratic error of prediction as it can be estimated directly from the data through a cross-validation procedure and because it has fewer assumptions regarding the distribution of horizon t and censoring indicator δ_i.

CHAPTER 4: ANALYSES

4.1 Simulations

The hazard regression (HARE) methodology within a joint latent class mixed model (JLCMM) framework was initially tested on simulated data. Data were constructed using a shared random-effects joint model with a Weibull parameterization. Survival times were estimated using the cumulative hazard inversion method (Bender, Augustin, & Blettner, 2005) where survival times are estimated using the equation $S_i^{-1}(u) = H_0^{-1}(-\log(u)\exp(-X_i^T\beta))$ with definitions $u \sim \text{Unif}(0,1)$, X is a vector of covariates, β are the corresponding parameter coefficients for X, and H_0 is the cumulative baseline hazard function. For a Weibull distribution, the survival times are estimated by

$$S_i^{-1}(u_i) = \left(\frac{-\log(u_i)}{\psi\exp(X_i^T\beta)}\right)^{1/\omega}$$

where $\psi > 0$ is the scale parameter and $\omega > 0$ is the shape parameter. In order to simulate the latent classes, three different Weibull parameterizations were used to construct three latent classes of different sizes and survival distributions. These three classes were assimilated into a single simulated data set where the joint distribution of the longitudinal outcome and time-to-event outcome followed a mixture distribution of Weibull distributions.

The longitudinal model was simulated for individual i at visit s as

$$y_i(s) = (\tilde{\beta}_{0k} + b_{i0}) + (\tilde{\beta}_{1k} + b_{i1})s + \tilde{\beta}_{2k}x_{i1} + \tilde{\beta}_3 x_{i2} + \tilde{\beta}_4 x_{i3} + \epsilon_i(s),$$

and the time-to-event model was simulated as

$$\lambda_i(t) = \lambda_{0k}(t) \exp\{\tilde{\theta}_{1k} x_{i1} + \tilde{\theta}_{2k} x_{i2} + \tilde{\theta}_{3k} x_{i3}\}$$

where $y_i(s)$ is the longitudinal outcome, $\tilde{\boldsymbol{\beta}}_{\boldsymbol{k}} = (\tilde{\beta}_{0k}, \tilde{\beta}_{1k}, \tilde{\beta}_{2k})$ is the vector of class-specific fixed-effects parameters, $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_3, \tilde{\beta}_4)$ is the vector of sample-level fixed-effects, $\mathbf{b} = (b_0, b_1)$ is the vector of random-effects parameters, $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_{1k}, \tilde{\theta}_{2k}, \tilde{\theta}_{3k})$ is the vector of survival parameters, and $\epsilon \sim \mathcal{N}(0, 0.1)$ is the random error. Variable x_1 is a mixture of uniformly distributed random variables where 25% of the sample is simulated from a range of 30 to 65, 35% of the sample is simulated from a range of 65 to 75, and 40% of the sample is simulated from a range of 75 to 85. After simulation, the x_1 values were scaled by centering at 70 and dividing by 10. Variable x_2 was simulated as a Bernoulli random variable ($x_2 \in \{0, 1\}$) with $p = \frac{1}{2}$. Variable x_3 was also a Bernoulli random variable ($x_3 \in \{0, 1\}$) with $p = \frac{1}{2}$. For the sake of simplicity, consider x_1 as a scaled age variable, x_2 as a treatment variable, and x_3 as a sex variable.

The random effects were simulated as bivariate normal with $b_i \sim N_2(0, \Sigma)$ where

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \rho\sigma_{11}\sigma_{22} \\ \rho\sigma_{11}\sigma_{22} & \sigma_{22}^2 \end{bmatrix}$$

and $\sigma_{11} = 0.4, \sigma_{22} = 0.2, \rho = 0.4$. Survival time t_i was simulated using the aforementioned cumulative hazard inversion method

$$t_i = S_i(u_i)^{-1} = \left(\frac{-\log(u_i)}{\psi_k \exp(\tilde{\theta}_{1k} x_{i1} + \tilde{\theta}_{2k} x_{i2} + \tilde{\theta}_{3k} x_{i3} + \alpha\mu_i(s))} \right)^{1/\omega_k},$$

where $\mu_i(s) = y_i(s) - \epsilon_i(s)$, and α is an association parameter that establishes the relationship between the longitudinal and survival outcomes. For all simulations, $\alpha = 0.05$. Variable $u_i \sim Unif(0, 1)$, and ψ_k, ω_k are the scale and shape parameters for the Weibull distribution of latent class k. Times ranged from 0 to 2 for the survival model where any value over 2 was set to

Table 4.1.1 Details of the Simulation

Parameters	Latent Class 1	Latent Class 2	Latent Class 3
n	200	150	50
$\tilde{\beta}_0$	1.500	0.500	3.000
$\tilde{\beta}_1$	0.500	-0.500	1.000
$\tilde{\beta}_2$	0.250	-0.250	0.125
$\tilde{\beta}_3$	0.350	0.350	0.350
$\tilde{\beta}_4$	0.500	0.500	0.500
$\tilde{\theta}_1$	0.250	0.550	-0.350
$\tilde{\theta}_2$	1.125	-0.050	0.350
$\tilde{\theta}_3$	0	0	0
Weibull shape (ω)	0.500	1.500	3.000
Weibull scale (ψ)	1.000	0.900	1.100

2.1. The longitudinal outcome was estimated within this time range with intervals of 0.5 between observations. These longitudinal times are listed in figures as visits 1 through 5. Individuals had different numbers of repeated measures corresponding to censoring time. Individuals who were observed for all five visits without experiencing the event were considered censored at time 2.1. To assess potential model mispecification and overfitting from HARE's algorithm, $\tilde{\theta}_3 = 0$ for all three latent classes, effectively making the hazard function uninfluenced by sex. $\tilde{\beta}_3$ and $\tilde{\beta}_4$ were held constant across classes. The total sample comprised $N = 400$ with $n = 200$ for latent class 1, $n = 150$ for latent class 2, and $n = 50$ for latent class 3. Table 4.1.1 provides details about the simulated joint outcomes for each latent class. Figure 4.1.1 illustrates the distribution of times-to-event and longitudinal trajectories for each latent class.

HARE was used to estimate the survival time knots and any potential covariates with potential covariate knots for the time-to-event model for the entire sample of $N = 400$ observations. This estimation used the `hare()` function from the **polspline** package (version 1.1.19) in R v4.1.0 (Kooperberg, 2020). Survival times and event indicators where 0 represented censoring and 1 represented event occurrence were provided to the `hare()` function while controlling for age, treatment, and sex. HARE returned a survival function that consisted of knots for times 0.001761498, 0.00462919, 0.01013831, 0.3350078, and 0.7397903 as well as for linear

estimates of age and treatment. HARE did not indicate that sex was important an important covariate for the survival model, which followed from the $\tilde{\theta}_3 = 0$ coefficient specified in the simulation. From this output, the survival model for the JLCMM was determined to incorporate age, treatment, and the aforementioned knots in time.

The JLCMM was estimated using the `Jointlcmm()` function from the **lcmm** package (version 1.9.3) in R v4.1.0 (Proust-Lima, et al., 2017). Both Weibull and HARE-specified survival models were tested and compared. The presupposition made for these comparisons was that a three-class Weibull model would provide the best fit. The assessment would determine whether the three-class HARE model was the best fitting among all HARE models and whether this model provided comparable results to the three-class Weibull model.

The class-specific fixed- and random- longitudinal effects were held constant between the two sets of models.

$$y_i(s) = (\tilde{\beta}_{0k} + b_{i0}) + (\tilde{\beta}_{1k} + b_{i1})(\text{visit}) + \tilde{\beta}_{2k}(\text{age}) + \tilde{\beta}_3(\text{treatment}) + \tilde{\beta}_4(\text{sex}) + \epsilon_i(s)$$

The primary distinction between these two sets of models was the time-to-event specification. For the Weibull models, the hazard model controlled for age, treatment, and sex.

$$\lambda_{ik}(t) = \lambda_{0k}(t) \exp\{\tilde{\theta}_{1k}(\text{age}) + \tilde{\theta}_{2k}(\text{treatment}) + \tilde{\theta}_{3k}(\text{sex})\}$$
$$\lambda_{0k}(t) \sim Weib(\omega_k, \psi_k)$$

For the HARE models, the hazard model controlled for age and treatment, but not sex as it was not considered important for the survival model. Additionally, the hazard knots were included as part of the class-specific baseline hazard.

$$\lambda_{ik}(t) = \lambda_{0k}(t) \exp\{\tilde{\theta}_{1k}(\text{age}) + \tilde{\theta}_{2k}(\text{treatment})\}$$
$$\lambda_{0k}(t) = \exp\{\sum_{j=1}^{p} \beta_{jk} B_j\}$$

Note that β in the $\lambda_{0k}(t)$ for the HARE model are different from the $\tilde{\boldsymbol{\beta}}$ vector in the longitudinal model. A total number of 1 through 4 classes were tested for the Weibull and HARE models. In `Jointlcmm()` the `hazard` option was specified as "Weibull" for the Weibull models and "7-manual-splines" for the HARE models with `hazardnodes` equal to the knots in time determined from HARE. The `logscale` option was kept 'FALSE' for the Weibull models to help with convergence but kept 'TRUE' for the HARE models to exponentiate the log-hazard $\alpha(t|\mathbf{x})$ determined by HARE. The 1-class solution for the Weibull parameterization was used to initialize all multiple class Weibull models, and the 1-class solution for the HARE parameterization was used to initialize all multiple class HARE models. Sets of parameters for the multiple class models (i.e., $k = 2, 3, 4$) were randomly generated from the 1-class solutions using the `gridsearch()` function with 30 repetitions for each estimation and maximum of 15 iterations for each repetition. This `gridsearch()` method provides the ability to use multiple sets of initial parameters to ensure that convergence of the Levenberg-Marquardt algorithm reaches a global maximum (Proust-Lima, et al., 2017).

Model summaries and fit statistics are provided in Table 4.1.2. These summaries provide the log-likelihood value, BIC (Schwarz, 1978), number of parameters estimated, and proportion of the sample assigned to each latent class. BIC was used as the decision criterion for best fitting model, where the smallest BIC indicated the best fit. The JLCMM results from this simulation are presented in Table 4.1.3. Since the class-specific B-spline coefficient estimates are incomparable with the class-specific Weibull coefficient estimates, median survival times ($T_{.50}$) and mean survival times ($E(T)$) were estimated for the two models. Because latent classification suffers from label switching, latent class assignment was matched to the true latent class using a proportion method. This assignment method assumed that latent class assignment of at least 85% indicated the label of the corresponding true latent class. For example, if at least 170 individuals from true latent class 1 were assigned to latent class 2, it was assumed that "assigned" class 2 referred to true latent class 1. Table 4.1.3 labels latent classes by their true simulated latent class. All subsequent results tables follow this convention.

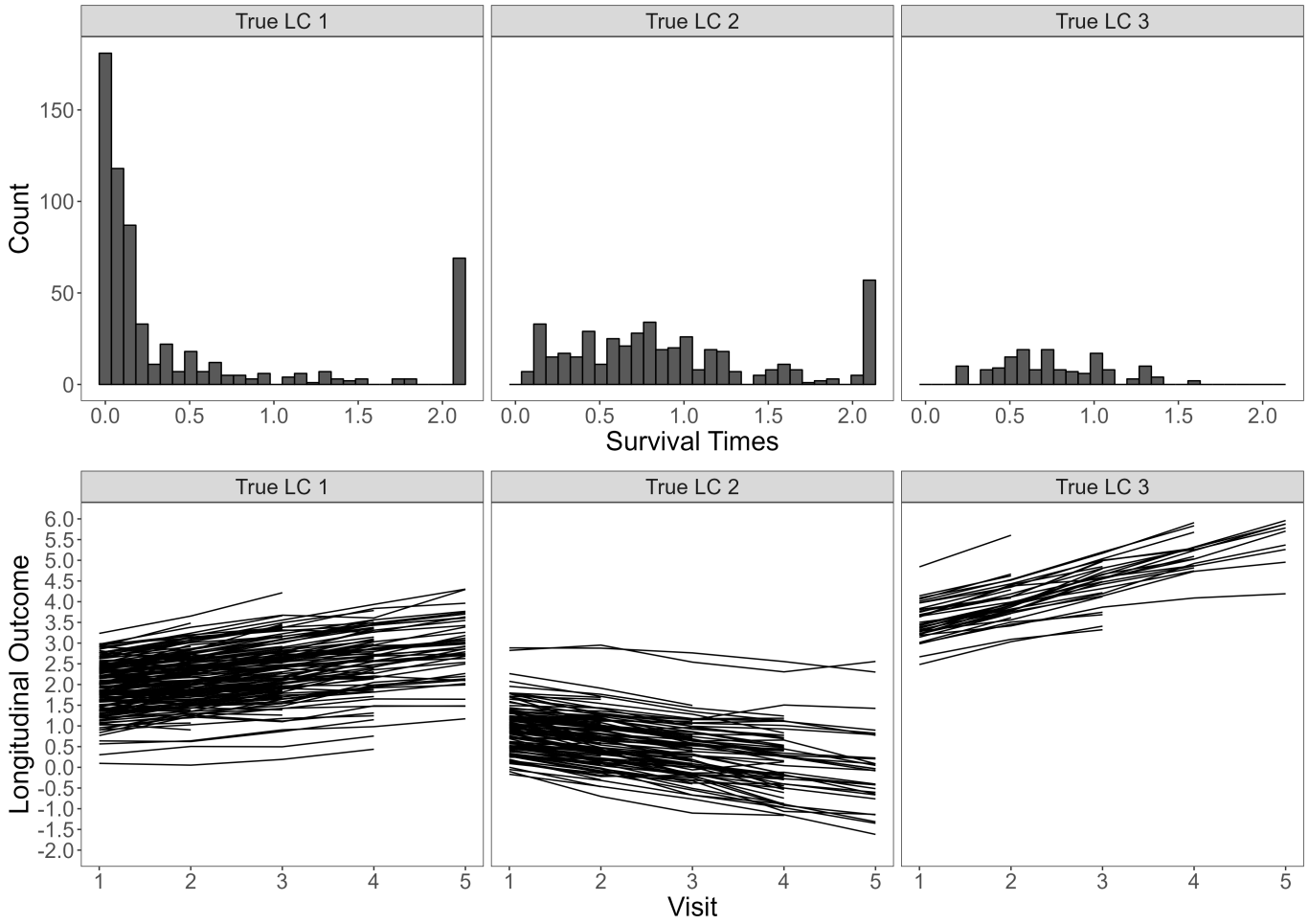

Figure 4.1.1 Histograms of Survival Times and Line Plots of Longitudinal Outcome for each Latent Class.

Table 4.1.2 Summary Table of Fit Statistics for the Eight JLCMM Models

Model	# parameters	$\ell\ell$	BIC	% LC 1	% LC 2	% LC 3	% LC 4
Weibull 1	14	-547.6	1179.1	100%	–	–	–
HARE 1	20	-518.8	1157.5	100%	–	–	–
Weibull 2	23	-397.7	933.2	63.00%	37.00%	–	–
HARE 2	35	-384.3	978.2	63.00%	37.00%	–	–
Weibull 3	32	-249.6	691.0	37.50%	50.75%	11.75%	–
HARE 3	50	-240.4	780.3	37.50%	50.50%	12.00%	–
Weibull 4	41	-246.5	738.7	37.50%	31.25%	19.00%	12.25%
HARE 4	65	-233.5	856.5	12.25%	37.75%	37.50 %	12.50%

We can see from the results in Table 4.1.2 that, unsurprisingly, the three class Weibull model has the best fit (BIC = 691.0) among all tested models. Among all HARE models, the 3 latent class solution also has the best fit (BIC = 780.3). While the three class Weibull model ought to have the best fit, note that the number of parameters in the three class HARE model (50 parameters) is much greater than that in the Weibull model (32 parameters), so some of the difference in BIC likely derives from some penalization in the number of parameters used in the HARE model. The proportions of each latent class for both three-class solutions were very similar (different by 0.25% at most), and both were very close to the true proportions (different by 0.75% at most). Both Weibull and HARE results overestimated the number of members in latent class 1 (LC 1) and underestimated the number of members in latent class 3 (LC 3). Both models correctly identified the number of members in latent class 2 (LC 2).

Table 4.1.3 summarizes the parameter estimates for the three-class Weibull and three-class HARE JLCMMs. Both the Weibull and HARE methods had minor bias for the longitudinal model estimates $\tilde{\boldsymbol{\beta}}_k$ and $\tilde{\boldsymbol{\beta}}$. In fact, the longitudinal model estimates for both Weibull and HARE methods were nearly identical in both the coefficient estimates and in the measurements of uncertainty. The bias in the survival models varied. The estimates for θ_{1k} were off by about 10-20% for both Weibull and HARE methods excepting for LC 3 where estimates were off by 70-80% with HARE having a slightly less biased estimate. Weibull and HARE estimated the intercept terms $\tilde{\beta}_{0k}$ relatively similarly, with $\tilde{\beta}_{02}$ being overestimated in both models. Similar levels of bias occurred for the θ_{2k} estimates for LC 1 and LC 3, but the estimate for LC 2 was

close to the simulated value. By default, the HARE method did not account for sex in the survival model. The Weibull method had somewhat large, yet non-significant estimates of the effect of sex $\tilde{\theta}_{3k}$ with LC 3 having the largest bias of 0.305.

Since survival times were simulated from Weibull distributions, $E(T)$ can be calculated directly from $E(T)_k = \psi_k \Gamma(1 + \frac{1}{\omega_k})$ where Γ is the gamma function $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$. Using this equation and the parameters from 4.1.1, $E(T)_1 = 2.0, E(T)_2 = 0.8125, E(T)_3 = 0.9822$. The estimated $T_{.50}$ for each latent class was estimated by finding the minimum time at which $S(T) \geq 0.50$ within the latent class. The estimated $E(T)$ was estimated by finding the area under the curve using the `AUC()` function from the **DescTools** package in R (Signorell et al., 2021). Note that these estimates were biased due to the presence of right censoring, so comparisons between models provided a better indication of the performance of HARE than comparison to the true values. While both methods produced incorrect estimates of mean and median survival time, the two methods had similar estimates to one another.

Table 4.1.3 Results of the Simulation

	Weibull			HARE		
Parameters	LC 1	LC 2	LC 3	LC 1	LC 2	LC 3
n	203	150	47	202	150	48
$\tilde{\beta}_0$	1.533 (0.042)	0.507 (0.047)	3.087 (0.071)	1.532 (0.042)	0.505 (0.047)	3.082 (0.071)
$\tilde{\beta}_1$	0.466 (0.018)	-0.526 (0.022)	1.028 (0.037)	0.466 (0.018)	-0.526 (0.022)	1.027 (0.037)
$\tilde{\beta}_2$	0.236 (0.021)	-0.199 (0.023)	0.112 (0.046)	0.235 (0.021)	-0.199 (0.024)	0.112 (0.046)
$\tilde{\beta}_3$	0.292 (0.042)	0.292 (0.042)	0.292 (0.042)	0.293 (0.042)	0.293 (0.042)	0.293 (0.042)
$\tilde{\beta}_4$	0.456 (0.041)	0.456 (0.041)	0.456 (0.041)	0.457 (0.042)	0.457 (0.042)	0.457 (0.042)
$\tilde{\theta}_1$	0.298 (0.072)	0.488 (0.082)	-0.641 (0.141)	0.301 (0.072)	0.487 (0.083)	-0.596 (0.133)
$\tilde{\theta}_2$	0.881 (0.160)	-0.034 (0.182)	0.691 (0.321)	0.879 (0.161)	-0.020 (0.181)	0.633 (0.313)
$\tilde{\theta}_3$	-0.196 (0.156)	0.207 (0.182)	0.305 (0.322)	0*	0*	0*
$T_{.50}$	0.298	0.848	0.912	0.318	0.827	0.848
$E(T)$	0.674	0.938	0.898	0.715	0.883	0.851

* = Parameter not estimated

Figure 4.1.2 shows the predicted longitudinal outcome, predicted hazard curve, and predicted survival curve for the three latent classes of the Weibull model (top) and HARE model (bottom). The predicted longitudinal outcomes between the two methods are practically identical, as expected from the results in Table 4.1.3. The shapes of the hazard curves for both methods match with HARE having higher maximum predicted hazards at the boundaries of

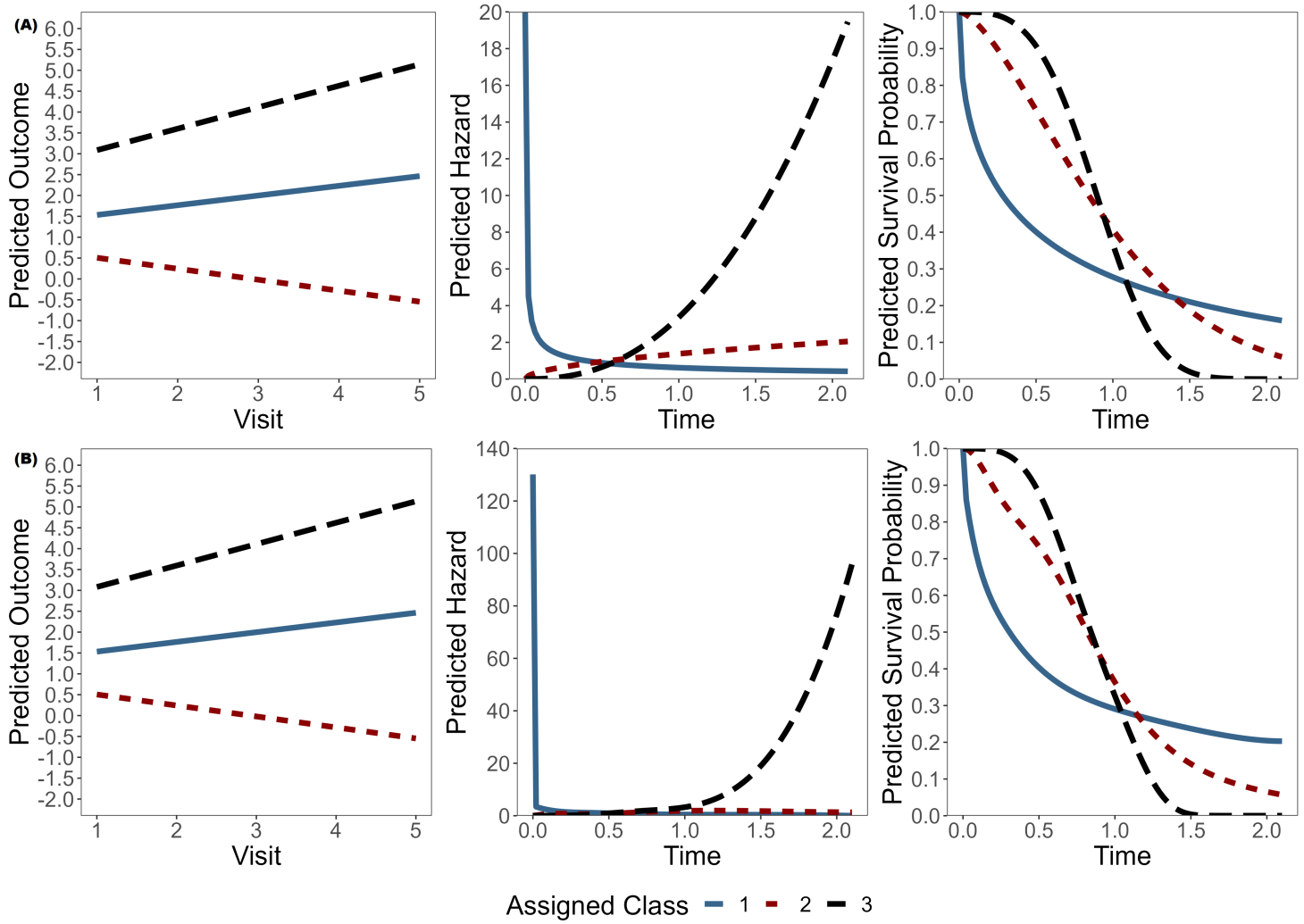

Figure 4.1.2 Predicted Longitudinal and Survival Outcomes from JLCMM for (A) Weibull and (B) HARE methodologies.

the survival times for LC 1 and LC 3. The smoothness of the Weibull hazard curve for LC 1 is greater than it is for HARE LC 1. Survival curves between the two methods also matched fairly well with some slight over estimation of survival probability among early time (0.5 - 0.7) within LC 2. Mean and median computation time for the JLCMM analyses were calculated using `microbenchmark()` from the R package **microbenchmark** (Mersmann, 2023). Mean computation time over 100 iterations for the JLCMM using a Weibull parameterization was 11.6 seconds (median = 11.4 seconds). Mean computation time over 100 iterations for the JLCMM using HARE was 738.6 seconds (median = 727.1).

Table 4.1.4 provides results from an analysis using simulated data of a sample of $N = 800$ where each latent class increased two-fold ($n_1 = 400, n_2 = 300, n_3 = 100$). Likewise, Table 4.1.5 uses simulated data from a sample of $N = 1200$ where each latent class increased three-fold from the original ($n_1 = 600, n_2 = 450, n_3 = 150$). For $N = 800$, estimates were overall less biased than for $N = 400$ with the exception of the estimates for LC 2. Both methods underestimated the effect of $\tilde{\beta}_{02}$, but these estimates also had lower standard errors than the simulation using $N = 400$. The longitudinal model tended to be less biased than the survival model for both methods, and that likely arises from the presence of censoring. Notably, both the Weibull and HARE have similar estimates for the survival parameters when the data were simulated according to a three-class Weibull mixture distribution. The Weibull model still erroneously estimated an effect for sex in the survival model, though these estimates were also closer to 0. The HARE model did not consider sex to be an important variable in the survival model for a sample of $N = 800$. Standard errors decreased for each estimate, indicating some asymptotically efficient behavior. For $N = 1200$, the HARE model in general produced the least biased results along with the smallest standard errors among all other HARE simulations. The HARE and Weibull models shared many similarities in the longitudinal and survival parameter estimates for these simulated data. Note that sex once again was not considered important by HARE even with a much alrger sample size, whereas the Weibull model estimated one non-zero

sex effect in LC 3 (albeit non-significant). These series of simulations indicated that HARE has some asymptotically favorable properties when used in JLCMM.

Table 4.1.4 Results of the Simulation for N = 800

	Weibull			HARE		
Parameters	LC 1	LC 2	LC 3	LC 1	LC 2	LC 3
n	392	307	101	393	306	101
$\tilde{\beta}_0$	1.494 (0.031)	0.482 (0.032)	3.007 (0.048)	1.493 (0.031)	0.481 (0.032)	3.006 (0.048)
$\tilde{\beta}_1$	0.508 (0.014)	-0.498 (0.016)	1.002 (0.027)	0.507 (0.013)	-0.499 (0.016)	1.002 (0.027)
$\tilde{\beta}_2$	0.245 (0.017)	-0.255 (0.019)	0.110 (0.033)	0.245 (0.017)	-0.225 (0.019)	0.110 (0.033)
$\tilde{\beta}_3$	0.338 (0.030)	0.338 (0.030)	0.338 (0.030)	0.339 (0.030)	0.339 (0.030)	0.339 (0.030)
$\tilde{\beta}_4$	0.568 (0.031)	0.568 (0.031)	0.568 (0.031)	0.569 (0.030)	0.569 (0.030)	0.569 (0.030)
$\tilde{\theta}_1$	0.285 (0.047)	0.555 (0.067)	-0.564 (0.101)	0.285 (0.048)	0.570 (0.068)	-0.571 (0.095)
$\tilde{\theta}_2$	1.119 (0.118)	-0.206 (0.127)	0.500 (0.209)	1.132 (0.122)	-0.223 (0.128)	0.479 (0.210)
$\tilde{\theta}_3$	-0.105 (0.109)	-0.047 (0.126)	-0.054 (0.228)	0*	0*	0*
$T_{.50}$	0.339	0.827	0.785	0.361	0.848	0.806
$E(T)$	0.715	0.914	0.790	0.775	0.932	0.795

* = Parameter not estimated

Table 4.1.5 Results of the Simulation for N = 1200

	Weibull			HARE		
Parameters	LC 1	LC 2	LC 3	LC 1	LC 2	LC 3
n	596	457	147	596	454	150
$\tilde{\beta}_0$	1.519 (0.025)	0.531 (0.026)	2.987 (0.041)	1.517 (0.025)	0.531 (0.026)	2.979 (0.041)
$\tilde{\beta}_1$	0.513 (0.012)	-0.492 (0.013)	1.022 (0.024)	0.513 (0.012)	-0.492 (0.013)	1.018 (0.024)
$\tilde{\beta}_2$	0.243 (0.021)	-0.237 (0.015)	0.122 (0.027)	0.243 (0.013)	-0.236 (0.015)	0.121 (0.026)
$\tilde{\beta}_3$	0.359 (0.024)	0.359 (0.024)	0.359 (0.024)	0.357 (0.024)	0.357 (0.024)	0.357 (0.024)
$\tilde{\beta}_4$	0.495 (0.024)	0.495 (0.024)	0.495 (0.024)	0.498 (0.024)	0.498 (0.024)	0.498 (0.024)
$\tilde{\theta}_1$	0.283 (0.039)	0.566 (0.054)	-0.318 (0.067)	0.279 (0.039)	0.573 (0.056)	-0.316 (0.068)
$\tilde{\theta}_2$	1.066 (0.098)	-0.018 (0.104)	0.490 (0.183)	1.066 (0.098)	-0.018 (0.104)	0.489 (0.183)
$\tilde{\theta}_3$	0.091 (0.091)	-0.040 (0.104)	0.214 (0.172)	0*	0*	0*
$T_{.50}$	0.424	0.827	0.870	0.806	0.848	0.361
$E(T)$	0.777	0.906	0.852	0.795	0.932	0.774

* = Parameter not estimated

One hundred replications were done for the $N = 400$ case using $m = 100$ different simulated data sets. The focus of these replications was to determine estimation properties of the HARE method for a three-class model, so model fit statistics were not computed, and Weibull models were not used as comparisons. Table 4.1.6 provides a summary of these replicates. Bias, empirical standard error (EmpSE), mean squared error (MSE), and coverage probability were provided in the table for all model parameter estimates. Most of these estimates show a

coverage of the true parameter value close to 95%. The primary exception was for the estimate of $\tilde{\theta}_{2k}$, which consistently had a 92% coverage probability for the three latent classes. This lower coverage probability may have arisen due to the estimate for LC 2 being so close to 0. A statistically significant estimate of a parameter close to 0 requires a larger sample size than one whose distance from 0 is large. Bias for certain estimates are greater than others, with estimates for the survival model typically larger than for the longitudinal model. Similarly, the MSE for the survival parameters were greater than for the longitudinal parameters in part due to greater empirical standard error and greater bias.

Table 4.1.6 One Hundred Replicates of JLCMM of Simulated Data Using HARE

			LC 1					LC 2					LC 3		
Parameters	Truth	Bias	EmpSE	MSE	Coverage Probability	Truth	Bias	EmpSE	MSE	Coverage Probability	Truth	Bias	EmpSE	MSE	Coverage Probability
$\tilde{\beta}_0$	1.500	-0.0045	0.0463	0.0022	0.95	0.500	-0.0014	0.0451	0.0020	0.98	3.000	-0.0075	0.0746	0.0056	0.93
$\tilde{\beta}_1$	0.500	0.0023	0.0196	0.0004	0.94	-0.500	0.0002	0.0216	0.0005	0.95	1.000	-0.0039	0.0414	0.0017	0.95
$\tilde{\beta}_2$	0.250	-2.67×10^{-6}	0.0218	0.0005	0.97	-0.250	-0.0001	0.0266	0.0007	0.94	0.125	0.0006	0.0507	0.0026	0.96
$\tilde{\beta}_3$	0.350	0.0040	0.0449	0.0020	0.95	0.350	0.0040	0.0449	0.0020	0.95	0.350	0.0040	0.0449	0.0020	0.95
$\tilde{\beta}_4$	0.500	0.0006	0.0391	0.0015	0.94	0.500	0.0006	0.0391	0.0015	0.94	0.500	0.0006	0.0391	0.0015	0.94
$\tilde{\theta}_1$	0.250	0.0144	0.0704	0.0052	0.93	0.550	0.0024	0.0991	0.0098	0.96	-0.350	-0.0051	0.1266	0.0161	0.95
$\tilde{\theta}_2$	1.125	0.0063	0.1905	0.0402	0.92	-0.050	0.0132	0.1712	0.0295	0.92	0.350	0.0926	0.3088	0.1040	0.92

Simulation results indicated that HARE provides a reasonable estimate of the survival model within a JLCMM without having any parametric assumptions about the survival functions. The true number of latent classes was accurately determined with minimal membership misclassification. Parameter estimates converged with true values as evidenced with similarities between the Weibull and HARE JLCMM estimates. Further, the HARE method accurately excluded sex from the survival model whereas the Weibull model estimated the parameter as being non-trivial in certain simulations. The ability for HARE to use data-driven methods to estimate the survival model provides support for its use when parametric assumptions are dubious, particularly in heterogeneous cases. If a joint model is homogeneous and parametric in nature, HARE provides comparable estimates to the true values. However, a parametric method would be unlikely to provide comparable estimates to HARE in a heterogeneous sample.

One potential limitation from using HARE is its overestimation of the hazard for boundary cases. The hazards for LC 1 early in time and LC 3 late in time were overestimated around the support of the knots. However, this overestimation did not affect the survival estimates for these

classes. Future work should be done such that knot selection may account for class-specific data support. Specifically, latent classes with fewer data around knots could have weighted estimates that mitigate overfitting.

4.2 A Joint Model of Mini-Mental State Examination Response and Time-to-Dementia Diagnosis

Data Description

The Paquid research data comprises a representative random sample of people aged 65 and older living in two regions of southwestern France, Gironde and Dordogne (Letenneur, et al., 1994). The goal of the Paquid study was to estimate the incidence of dementia, vascular dementia, and Alzheimer's disease among older adults living in France. The data were collected as a prospective cohort with random sampling stratified by age, sex, and size of the geographical sampling unit (based on electoral rolls of 37 parishes). Each stratum contains participants who belonged to one of three age groups (65 - 74 years, 75 - 84 years, and $\geq$ 85 years), either male or female sex, and one of five geographic units ($<$ 2000 inhabitants, 2000 - 9999 inhabitants, 10000 - 49999 inhabitants, 50000 - 99999 inhabitants, and $\geq$ 100000 inhabitants). The final sample included $n = 4050$ at baseline. More details about the Paquid data sample can be read in Letenneur, et al. (1994).

The **lcmm** package for R (Proust-Lima, et al., 2017) contains a subset of these data comprising 2250 observations of 500 participants and 12 variables from the Paquid cohort. Data include repeated measures of cognitive assessments such as the Mini-Mental State Examination (MMSE) (Folstein, et al.; 1975), an indicator of dementia (1 = dementia, 0 = censored), and time to dementia (or time to censoring). Covariates of interest included age, sex, and an education variable (binary with 1 = graduated from primary school and 0 = otherwise), where sex and education were time-invariant. Data were collected from baseline over a period of a maximum of 20 years. The MMSE is a test of global cognitive functioning with integer scores ranging from

0 to 30, where a higher score indicates higher global cognitive function. Folstein, et al. (1975) developed the MMSE to assess participants quickly (typically in the span of 5 - 10 minutes) to accommodate participants with delerium or dementia. However, this instrument was normed using a sample including participants with dementia, affective disorders such as depression or depression with cognitive impairment, schizophrenia, mania, neurosis, and personality disorders with drug abuse. The participants with dementia on average scored the lowest ($\bar{x} = 12.2$) of all cognitive impairment disorders, and scores of 20 were found only in participants with dementia or some functional psychosis. Folstein, et al. (1975) report that the instrument is valid and reliable (including test-retest reliability) as a measure of cognitive impairment. As such, the MMSE can provide insight into the development of dementia and may have value as a longitudinal marker of cognitive functioning for this sample.

Methods

The Paquid study data were used to determine the feasibility of using the HARE approach to estimate the conditional log-hazard of the time-to-onset of dementia within a JLCMM. This model also estimated the change in MMSE scores as a joint measure of cognitive functioning. This score change was estimated using a linear mixed-effects model where random-effects accounted for the within-subject variance produced by the longitudinal nature of the data. This JLCMM accounted for the potential heterogeneity of these joint outcomes by estimating latent classes. This method was compared to the JLCMMs detailed in the `Jointlcmm()` function vignette of the **lcmm** package documentation (Proust-Lima, et al., 2017). Model specifications were kept nearly identical for the purposes of comparison. The distribution of the MMSE scores is non-normal, so a normed version of the MMSE derived from the **NormPsy** (Phillips, et al., 2014) package was used for the JLCMMs. Further, age was centered at 65-years-old and divided by 10 to mitigate estimation issues of using a quadratic term with large ages. Education and sex were not transformed. The longitudinal mixed-effects model estimating change in normed MMSE was specified as

$$Y_{ij} = (\beta_0 + b_0) + (\beta_{j1} + b_{j1})x_{ij1} + (\beta_{j2} + b_{j2})x_{ij1}^2 + \beta_{j3}x_{i2} + \epsilon_{ij}$$

where Y_{ij} is the normed MMSE score for participant i at time j, x_{ij1} is the normed age for participant i at time j, x_{i2} is the education status for participant i, β are the fixed-effects parameters, b are the random-effects parameters, and ϵ_{ij} are the error terms for each participant i at time j. The polynomials for the normed age variable were orthogonal and calculated using R's `poly()` function in R v 4.1.0 (R Core Team, 2021). Class-specific estimates were determined for the age polynomials, but class membership was estimated as an intercept-only model. The class-specific proportional hazards model was estimated as

$$\lambda_i(t)\Big|_{c_i=k} = \lambda_{0k}(t)e^{\theta_{i1}x_{i2}+\theta_{i2}x_{i3}}$$

for $k = 1, \ldots, K$ classes, θ_{i*} fixed-effects parameters in the survival model, x_{i2} as education status, x_{i3} as sex (female referent), and λ_{0k} as the class-specific baseline hazard function. Models with 1 to 4 potential latent classes were estimated.

The distinction between the Proust-Lima, et al. (2017) specification of the model and the HARE specification entails the estimation of $\lambda_i(t)$ where the former specifies a two-parameter Weibull distribution to estimate the log-hazard function and the latter estimates the log-hazard using the adaptive HARE technique. Spline-based specification of the log-hazard function was given in the `hazard` option of the `jlcmm()` function and its knots were specified using the `hazardnodes` option. Two HARE models were estimated: one where class-specific estimates were unrelated (`hazardtype = ''Specific''`) and one where class-specific estimates were proportional (`hazardtype = ''Proportional''`). The proportional method assumed that all class-specific estimates of the baseline hazard function were proportional, which allowed for the estimation of fewer parameters than the 'specific' specification. These two specifications were tested in the HARE model based on observations made from the simulation study. Since HARE models estimated a larger number of parameters compared to

a parametric survival models, it was of interest to determine whether some specifications to reduce the parameter space could improve model fit. Model comparisons were made with fit indices using the `lcmm::summarytable()` function. While log-likelihood and AIC were examined, decisions of class structure were made with minimizing the BIC while accounting for substantive reasoning for the latent class structure. After determining the best Weibull and HARE models, class-specific predicted trajectories of normed MMSE were compared between the methods. Cross-entropy was calculated to assess the predictive ability of latent classification across several ages (i.e., 70, 72, 75, 77, 80, 82, 85, 87, and 90) using the `epoce()` function from the **lcmm** package with profiles of cross-entropy plotted for comparison.

Results

Figure 4.2.1 provides an illustration of the empirical hazard rate and survival function over a nearly 20-year period (ages 65 to 85) for the Paquid study data. These empirical estimates were determined by treating the survival functions as counting processes. The survival function was calculated as $S(t) = I(T \geq t)$ where $I(.)$ is the indicator function, and the hazard rate was caculated as $\lambda(t)dt = P(t \leq T < t + dt, C = 1|T \geq t)$ where dt is set as an interval of 2 years and $C = 1$ refers to experiencing the event (i.e. dementia onset). While the hazard rate primarily increases over the study period, as reflected in the survival curve, there are a few intervals where this hazard rate peaks. These peaks occur within the first few years of study, around 10 years into the study, and between 15 and 20 years into the study. This latter interval refers to ages of 80 to 85, and its hazard rate is expectedly the largest during the study period.

Table 4.2.1 summarizes the empirical hazard rates and survival probabilities of this sample. Column N indicates the number not censored or with dementia at the beginning of the interval, column $Cases$ indicate the number of participants who were diagnosed with dementia within that interval, the hazard rate calculates the number of dementia cases divided by N by the end of an interval, and the left-interval survival probability estimates the number of participants who

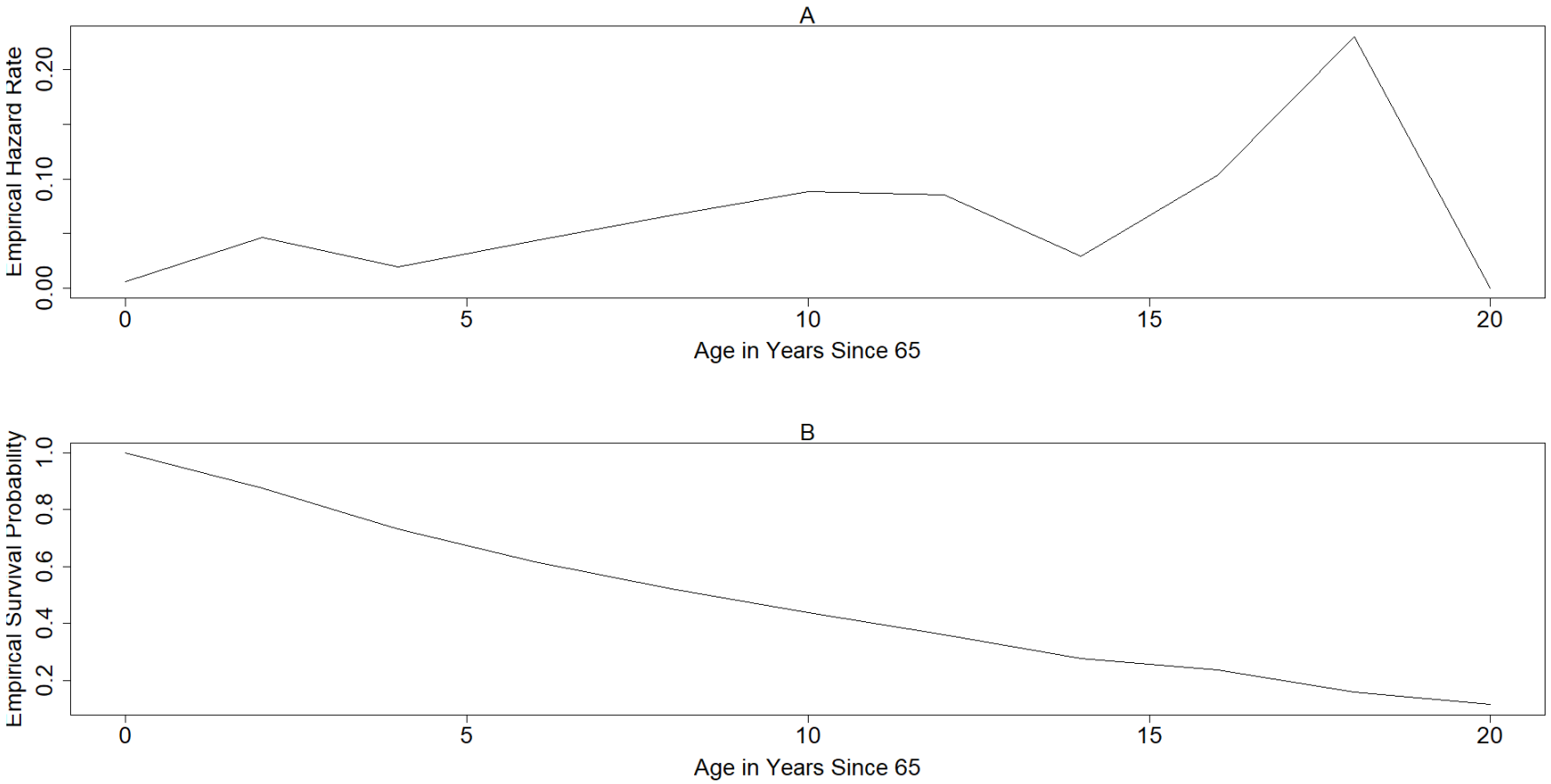

Figure 4.2.1 (A) Empirical Baseline Hazard Function of Paquid Data (B) Empirical Survival Function of Paquid Data

are dementia-free or been censored at the *beginning* of an interval. All recorded events after 20 years were right-censored.

Table 4.2.1 Empirical Survival Estimates of Paquid Data

Interval	Ages	N	Cases	Hazard Rate	Left-Interval Survival Probability
$[0, 2)$	65 - 67	490	3	0.006	1.000
$[2, 4)$	67 - 69	429	20	0.047	0.876
$[4, 6)$	69 - 71	359	7	0.019	0.733
$[6, 8)$	71 - 73	302	13	0.043	0.616
$[8, 10)$	73 - 75	255	17	0.067	0.520
$[10, 12)$	75 - 77	214	19	0.089	0.437
$[12, 14)$	77 - 79	176	15	0.085	0.359
$[14, 16)$	79 - 81	136	4	0.029	0.278
$[16, 18)$	81 - 83	116	12	0.103	0.237
$[18, 20)$	83 - 85	78	18	0.231	0.159
$[20, \infty)$	85+	56	0	0.000	0.114

Figure 4.2.2 illustrates a comparison of the estimated hazard function and survival curve from `Jointlcmm()` when using a one-class Weibull parameterization versus using HARE to estimate the baseline hazard. For clarity, the Weibull parameterization from **lcmm** differs from that specified in §4.2.1 by defining $\lambda_0(t) = \zeta_1\zeta_2 t^{\zeta_2 - 1}$, where ζ_1 is the scale parameter and ζ_2 is the shape parameter. The B-spline knots for the hazard function were estimated by HARE as 18.57, 18.90, and 19.7, which corresponded to ages 83.57, 83.9, and 84.7

Both the HARE and the Weibull models estimate an increasing hazard function (and, conversely, a decreasing survival). The primary difference, however, is that a Weibull parameterization estimates the hazard as a strictly increasing function whereas HARE estimates a peak around the 18-year followup mark. This distinction is also clear in the survival functions as the HARE estimated survival function contains an "elbow" near the 18-year mark where a sharp increase of dementia cases occur. The HARE-estimated hazard function reflects the maximum hazard around the 18-to-20 year follow-up as seen in Table 4.2.1 ($\lambda(t) = 0.231$) that the Weibull parameterization does not. This provides some evidence that data-driven methods for estimating the baseline hazard function such as done with HARE provides a more robust

estimate of the hazard function than a parametric estimate. The utility of B-splines to adapt to the data may lead to estimate heterogeneity within the sample better than a parametric method.

Table 4.2.2 Joint Latent Class Mixed Model Summaries of Standard Weibull & HARE models

Model	K	LL	npm	BIC	% Class 1	% Class 2	% Class 3	% Class 4
Weibull	1	-7747.9	15	15588.7	100%			
HARE	1	-7746.3	20	15616.5	100%			
Weibull	2	-7696.9	21	15553.9	70%	30%		
HARE PH[1]	2	-7690.9	25	15536.7	17.1%	82.9%		
HARE SH[2]	2	-7693.3	31	15578.7	78.0%	22.0%		
Weibull	3	-7681.4	27	15530.1	55.1%	21.8%	23.1%	
HARE PH	3	-7675.3	30	15536.4	21.0%	62.3%	16.7%	
HARE SH	3	-7680.6	43	15627.5	22.6%	59.8%	17.6%	
Weibull	4	-7678.4	33	15521.2	37.6%	27.8%	5.1%	29.6%
HARE PH	4	-7661.7	35	15540.1	16.5%	2.2%	62.0%	19.2%
HARE SH	4	-7677.9	55	15696.7	9.4%	33.9%	36.9%	19.8%

1 = "PH" is Proportional Hazards
2 = "SH" is Specific Hazards

Table 4.2.2 summarizes the model fit results assessing the single class model (as depicted in Figure 4.2.2) as well as for models with 2, 3, and 4 latent classes. The algorithm for estimating the two-, three-, and four-latent class models used starting values from the one-latent class model to estimate the parameters. Both Weibull parameterization and HARE estimation methods are listed. The negative log-likelihood and BIC are provided as fit statistics with model selection focused on BIC. The number of parameters estimated is listed under the "npm" column. The proportion of the sample classified in each latent class is given in the latter four columns. Different estimation methods were used for the models to ensure finding the global maximum of solutions. For the Weibull models, starting values were initialized with the `B` option in `Jointlcmm()` following the best fitting results found in Proust-Lima, et al. (2017). For the HARE models, best fits were found using the `gridsearch()` function within **lcmm** specifying 30 repetitions, 15 maximum iterations, and the one-class HARE model as the model from which initial values are generated. This function randomly generates a number of initial values based on the initial model equal to the number of repetitions specified (i.e., 30). The

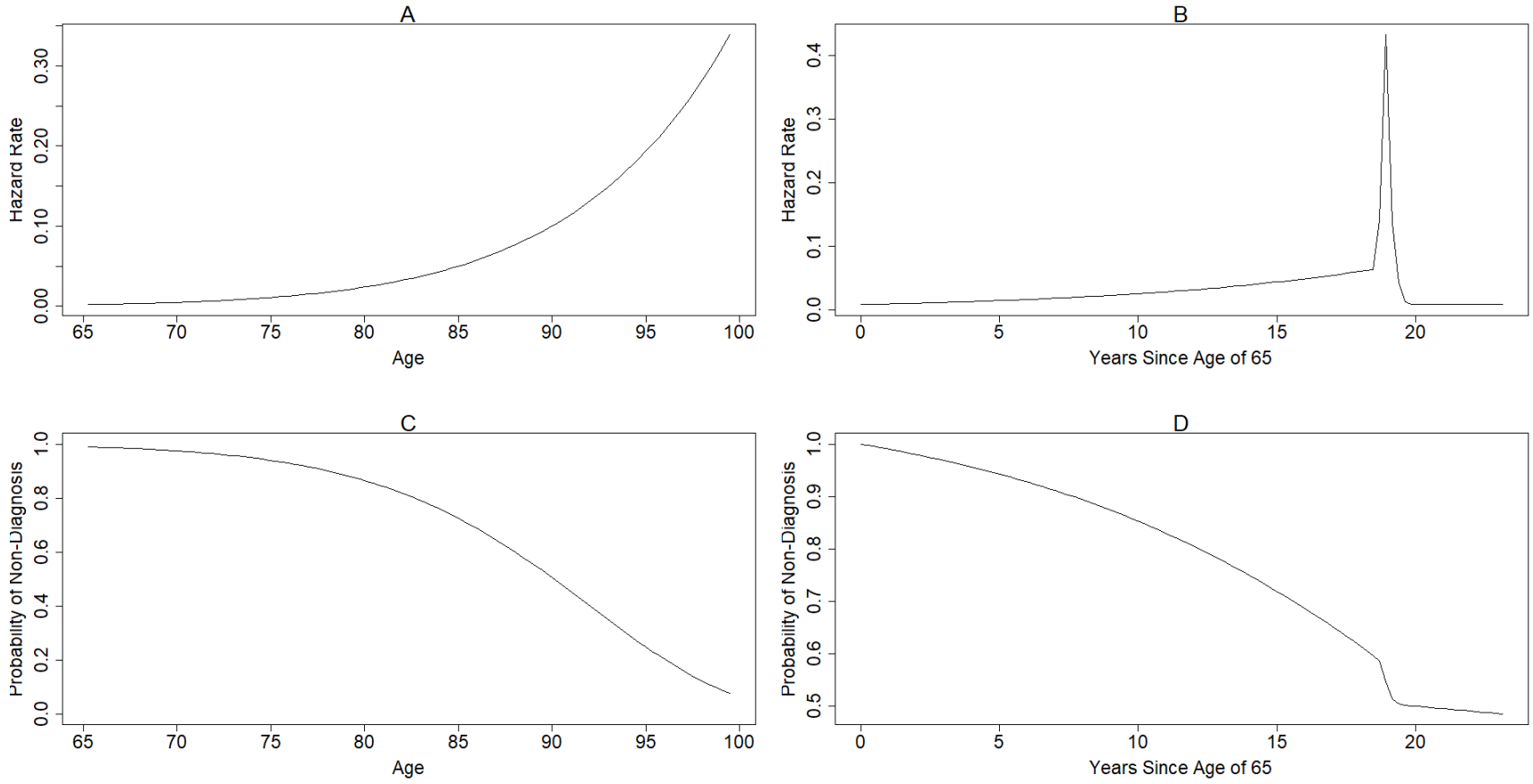

Figure 4.2.2 (A) Weibull estimated hazard function (B) HARE estimated hazard function (C) Weibull estimated survival curve (D) HARE estimated survival curve

function then finds the set of values which reaches the maximum likelihood over the specified number of iterations (i.e., 15). The log-likelihood is saved from these repetitions, and the algorithm then uses the best of all of these log-likelihoods to begin the final iteration. These parameters are used to find a solution for the Levenberg-Marquardt algorithm. If the convergence criteria are met, these estimates are the solution to the parameter estimation algorithm.

The best fitting model among those tested is the four-class Weibull model (BIC = 15521.2). The best fitting HARE model was the three-class PH model (BIC = 15536.4). While the four-class Weibull model did provide the best fit to the data, it is worth noting two points: (i) the Weibull model only provided best fit when providing a predetermined set of initial values that had been validated in previous literature, and (ii) the difference in BIC is 15.2, which is relatively negligible. The former point becomes more apparent when detailing that standard estimation and `gridsearch()` methods of estimating the global maximum provided worse fits than what was determined by HARE. It is possible that HARE could have a lower BIC if a set of optimal parameters were set with the `B` option. Since the PH models provided better fit over the SH models, and the proportion of latent class assignment was very similar between PH and SH models, only the three-class HARE PH model was compared with the four-class Weibull solution. For the three-class HARE PH solution, LC 2 comprises the majority of the sample with 62.3% belonging to this class. LCs 1 & 3 have comparable sizes (21% and 16.7%, respectively). For the four-class Weibull solution, most members were assigned to LCs 1, 2, and 4 (37.6%, 27.8%, and 29.6%, respectively) and 5.1% for latent class 3.

Tables 4.2.3, 4.2.4, 4.2.5, and 4.2.6 summarize the latent classification through posterior probability for the four-class Weibull model. Table 4.2.3 gives information that parallels the information given in Table 4.2.2. Table 4.2.6 summarizes the same latent class assignment if *only* the longitudinal normalized MMSE model were used to estimate the posterior probability of being a member of a latent class. Comparing tables 4.2.3 and 4.2.6 can provide insight to how the joint model (i.e., including time-to-event) affects latent class assignment. If the time-to-dementia onset were ignored from this model, then class 3 would be much smaller

comprising $< 1\%$ of the whole sample, and latent class 4 would be much larger. Table 4.2.4 provides the mean class assignment probabilities within each class. Since class membership is assigned modally, each participant has a posterior probability of being in any of the four classes. A participant becomes assigned to a specific latent class if that participant's probability of being a member in that class was the highest. As an example, for $c_i = 2$, the mean posterior probability of being assigned to class 2 is 0.598 or around 60%. This table provides some insight into counterfactual inference. For participants assigned to class 1, the second most likely class membership is class 2 where $P(c_i = 2) = 0.205$. However, participants assigned to class 1 have a low probability, on average, of being assigned to classes 3 or 4 with probabilities around 6%. Table 4.2.5 illustrates how strong these posterior probabilities are within each class assignment. The least uncertain class assignments were in latent classes 1 & 4, who had around half of their posterior probabilities ≥ 0.70. Further, over 40% of the posterior probabilities of those assigned to class 4 were ≥ 0.90. Assignment to classes 2 or 3 contained much more uncertainty, with less than 40% of their respective members having posterior probabilities ≥ 0.70. Class 3 had only 8% of its posterior probabilities ≥ 0.90.

Table 4.2.3 Posterior Probability of Class Assignment for Best Fitting Weibull Model

	Class 1	Class 2	Class 3	Class 4
N	184	136	25	145
%	37.6	27.8	5.1	29.6

Table 4.2.4 Mean of Posterior Probabilities for Best Fitting Weibull Model

k	$P(c_i = 1)$	$P(c_i = 2)$	$P(c_i = 3)$	$P(c_i = 4)$
1	0.669	0.205	0.063	0.064
2	0.172	0.598	0.121	0.109
3	0.145	0.159	0.652	0.044
4	0.087	0.159	0.069	0.685

Table 4.2.5 Proportion of Posterior Probabilities above Thresholds for Best Fitting Weibull Model

	k = 1	k = 2	k = 3	k = 4
$P(c_i = k) \geq 0.7$	49.5%	31.6%	36.0%	53.8%
$P(c_i = k) \geq 0.8$	34.8%	29.4%	24.0%	47.6%
$P(c_i = k) \geq 0.9$	25.5%	23.5%	8.0%	40.7%

Table 4.2.6 Posterior Probability of Class Assignment for Best Fitting Weibull Model using only longitudinal model

	Class 1	Class 2	Class 3	Class 4
N	186	125	4	175
%	38.0	25.5	0.8	35.7

Tables 4.2.7, 4.2.8, 4.2.9, and 4.2.10 summarize the latent classification through posterior probability for the three-class HARE model. Table 4.2.7 gives information that parallels the information given in table 4.2.3. Table 4.2.10 summarizes the same latent class assignment if *only* the longitudinal normalized MMSE model were used to estimate the posterior probability of being a member of a latent class. Like with the Weibull model, we can compare tables 4.2.7 and 4.2.10 to provide insight to how the joint model (i.e., including time-to-event) affects latent class assignment. If the time-to-dementia onset were ignored from this model, then class 3 and class 1 would have fewer and more assignments than otherwise, respectively. Class 2 would largely be unchanged. Table 4.2.8 provides the mean class assignment probabilities within each class. With fewer total classes in this model than the previous Weibull model, a direct comparison is inappropriate. However, the mean posterior probabilities of these three classes are higher overall indicating less uncertainty with class assignment. Class 1 has the largest mean posterior probability, and two-thirds of its members have posterior probabilities ≥ 0.90 (Table 4.2.9). Classes 2 and 3 had converse rank-order. That is, the most probable assignment for class 2 was class 2 and the second-most probable assignment was class 3. Conversely, the most probable assignment for class 3 was class 3, and the second most probable was class 2. Neither class 2 nor 3 had a high mean posterior probability of assignment to class 1.

Posterior probabilities and class assignments for the three-class HARE model were overall higher and more modal than the four-class Weibull model. Part of this distinction derives from the HARE model estimating fewer latent classes than the Weibull model. However, there is still clearer discrimination between classes in the HARE model than the Weibull model as can be seen comparing tables 4.2.4 & 4.2.5 with tables 4.2.8 & 4.2.9. A greater proportion of class assignments reach the 0.7, 0.8, and 0.9 thresholds in the HARE model than the Weibull model with only the fourth latent class in the Weibull model having comparable posterior probabilities to the ones estimated in the HARE model.

Table 4.2.7 Posterior Probability of Class Assignment for Best Fitting HARE Model

	Class 1	Class 2	Class 3
N	103	305	82
%	21.0	62.3	16.7

Table 4.2.8 Mean of Posterior Probabilities for Best Fitting HARE Model

k	$P(c_i = 1)$	$P(c_i = 2)$	$P(c_i = 3)$
1	0.831	0.067	0.102
2	0.101	0.712	0.187
3	0.080	0.203	0.712

Table 4.2.9 Proportion of Posterior Probabilities above Thresholds for Best Fitting HARE Model

	k = 1	k = 2	k = 3
$P(c_i = k) \geq 0.7$	74.8%	51.2%	50.0%
$P(c_i = k) \geq 0.8$	70.9%	40.3%	42.7%
$P(c_i = k) \geq 0.9$	66.0%	30.5%	34.2%

Table 4.2.10 Posterior Probability of Class Assignment for Best Fitting HARE Model using only longitudinal model

	Class 1	Class 2	Class 3
N	120	300	70
%	24.5	61.2	14.3

Figure 4.2.3 illustrates the trajectory of outcomes for each class within the four-class Weibull model and three-class HARE model. The left figures illustrate the mean scores within each latent class over the study period. The right figures illustrate the probability of being dementia-free within each latent class over the study period. The Weibull model has distinct MMSE mean trajectories and dementia-free probabilities for the first (solid black line) and fourth (dotted blue line) latent classes. Cognitive functioning was consistently high for the first latent class, which reflects a high probability of being dementia-free with only some drop occurring in the patients' early-90s. Cognitive functioning for the fourth latent class was almost equal to the first class at baseline, but it decreased by approximately 20 points within 2 decades. Likewise, the probability of remaining dementia-free for this latent class diminished, with a decrease in probability occurring shortly after baseline and reducing to 0 by age 87. Latent class 2 (red dotted line) has a cognitive functioning score that decreases roughly twenty points, but its rate of decline occurs at a slower rate than latent class 4–a twenty point decrease over almost 3 decades. Its dementia-free probability decreases at a rate between that of classes 1 and 4. Overall, these three classes behave like a gradient with latent class 1 having better outcomes than latent class 2, which has better outcomes than latent class 4. The standout latent class is latent class 3, which has poor MMSE outcomes and good dementia-free probability outcomes. The third latent class starts with the highest mean cognitive functioning level, but it decreases almost 40 points within 3 decades. However, its probability of the onset of dementia is comparable to latent class 4. This latent class appears to have an elongated outcome of time spent dementia-free in spite of cognitive decline. One consideration is that this latent class is the smallest of the four (n = 25), which could account for the high MMSE average at baseline.

The HARE model illustrates mean MMSE trajectories and dementia-free probabilities that parallel the latent class outcomes for the Weibull model. Latent class comparisons can be seen in Table 4.2.11. Specifically, HARE latent class 1 (solid black line) is composed of members entirely from Weibull latent class 4. HARE latent class 2 (dotted red line) is primarily composed of members from Weibull latent class 1, but also has members from latent classes 2 -

4. The HARE latent class 3 (green dotted line) comprises mostly members of Weibull latent classes 2 with a few from the other latent classes. This blending of latent classes naturally raises the question whether the Weibull model erroneously separated one class into two or if the HARE model neglected to separate a naturally occurring group of people who decline in cognitive function without onset of dementia. The answer to this question likely depends on the substantive question itself (Bauer, 2022). For example, it may be integral to identify a latent class of people who do not qualify for a diagnosis of dementia in spite of decrease in cognitive function. In this case, the Weibull model may have superior qualities compared to this HARE model, especially since latent class 2 for HARE appears to contain members who identify as having a positive prognosis of dementia and cognitive functioning. However, considering latent class 3 in the Weibull model comprises a total of 25 people, the Weibull model may also be biased by a few outliers. This conjecture has some evidence by virtue of Weibull class 3 being primarily members of HARE class 2 with only a few members (n = 6) in HARE class 3. One conclusion from these figures is that the HARE model does not estimate the joint outcomes for its latent classes much different from the published model in Proust-Lima, et al. (2017).

Table 4.2.11 Latent Class Assignment Between the Two Models

HARE \ Weibull	LC 1	LC 2	LC 3	LC 4
LC 1	0	0	0	103
LC 2	181	65	19	40
LC 3	3	71	6	2

Figure 4.2.4 illustrates the predicted normed MMSE scores for each latent class within the Weibull model (top) and HARE model (bottom). These figures parallel figure 4.2.3 in that Weibull latent class 1 and HARE latent class 2 have better cognitive functioning scores over the study period compared to their respective other latent classes. The one distinction worth noting is that the Weibull model produces less discrimination between the predicted MMSE scores of

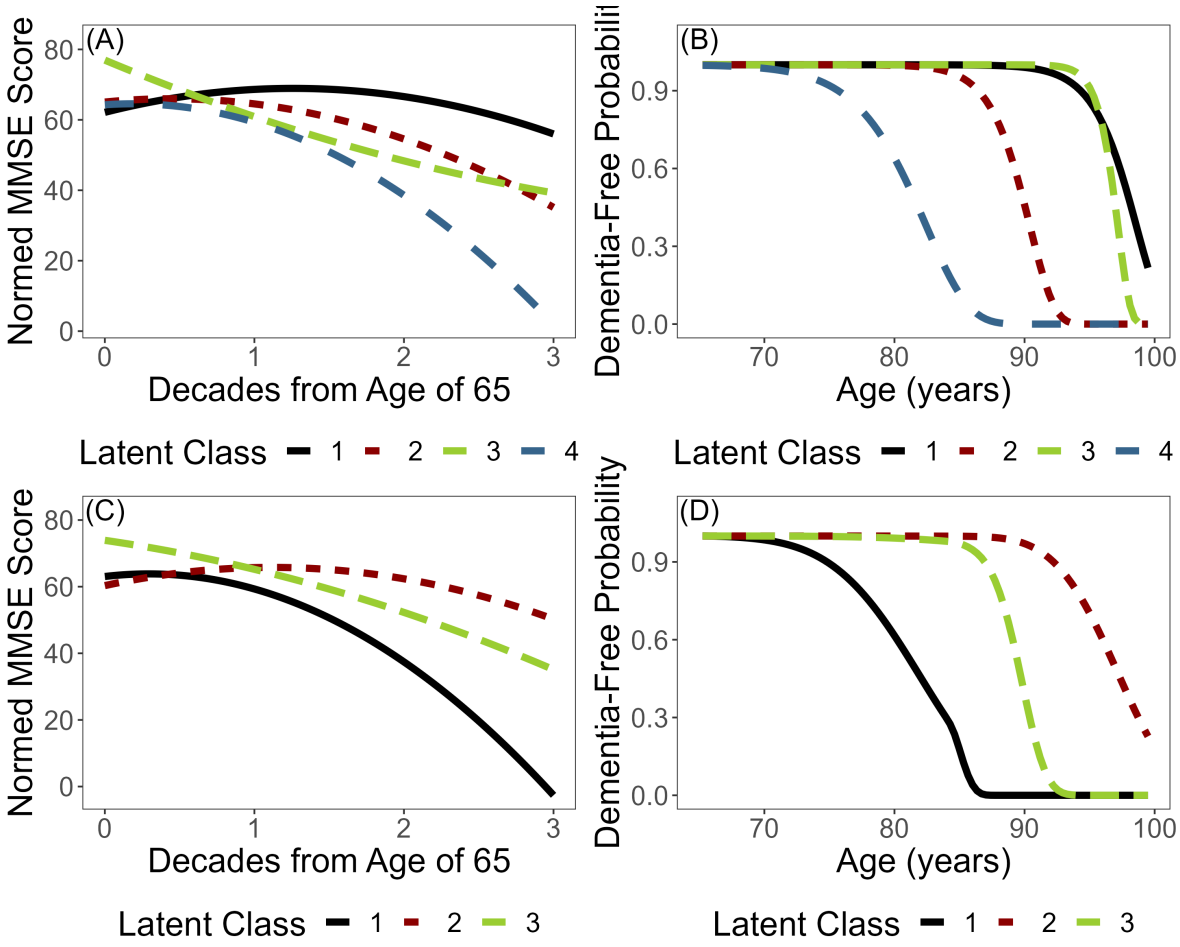

Figure 4.2.3 (A) Mean Trajectory of Normed MMSE for four-class solution (Weibull) (B) Class-specific Probabilities of Being Dementia-Free (Weibull)
(C) Mean Trajectory of Normed MMSE for three-class solution (HARE) (D) Class-specific Probabilities of Being Dementia-Free (HARE)

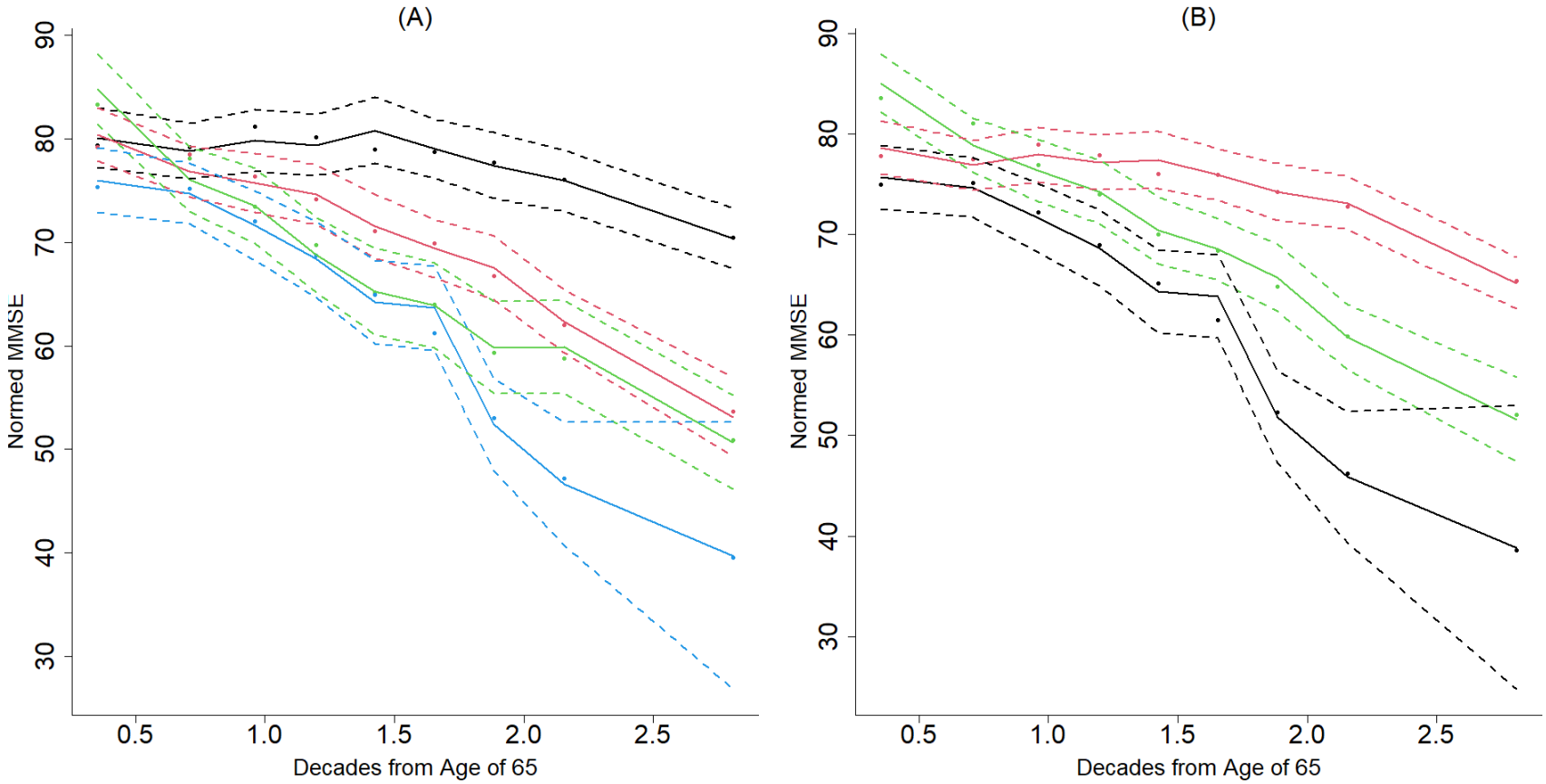

Figure 4.2.4 (A) Predicted Normed MMSE for four-class solution (Weibull) (B) Predicted Normed MMSE for three-class solution (HARE)

classes 2-4 later in the study period than the HARE model does between latent classes 1 and 3. Table 4.2.12 provides more details about this discrimination at later study periods.

Table 4.2.12 summarizes the expected prognostic observed cross-entropy (EPOCE) of the joint models (Commenges, et al., 2012). Specifically, the cross-validated prognostic observed log-likelihood (CVPOL) is calculated for each model and compared at the different ages. The CVPOL is a risk function that assess the predictive accuracy of the two outcomes in a joint model by calculating the expected cross-entropy under a leave-one-out cross-validation method and taking the average of these estimates. Lower CVPOL indicates a better prediction of the 'prognosis' of the joint model: the estimated normed MMSE and onset of dementia at each prediction time.

Table 4.2.12 EPOCE of The Weibull and HARE models

Age	N at risk	N events	CVPOL Weibull	CVPOL HARE
70	477	128	1.114	1.191
72	456	126	1.133	1.301
75	433	125	1.170	1.284
77	405	122	1.195	1.365
80	347	107	1.173	1.318
82	305	96	1.171	1.313
85	237	73	1.131	1.173
87	181	51	1.142	1.000
90	91	28	1.305	0.824

Overall, the four-class Weibull model has a lower CVPOL value for predictions at earlier ages. However, the three-class HARE model has a much lower CVPOL value at the 87-year-old and 90-year-old prediction times. These cross-entropy values reflect what is seen in Figures 4.2.2 and 4.2.4: the precise estimation of the hazard function from HARE's B-spline procedure allows for better prediction around the knots. This result should be no surprise as it is a semi-parametric approach. The jump in the hazard function, as seen in Figure 4.2.1, is not easily estimated using a parameterization such as the Weibull model–even when the Weibull parameters fit the data very well. One improvement here would allow for HARE to be more flexible within latent

classes, but first an examination of the model results is needed. Tables 4.2.13 & 4.2.14 provide the joint model estimates for the Weibull and HARE models, respectively.

The results of Tables 4.2.13 & 4.2.14 provide summaries for the four parts of a joint LCMM: class-membership estimates, the survival model estimates, the fixed-effects estimates of the longitudinal model, and the random-effects estimates of the longitudinal model (denoted Σ in the table). The class-membership estimates calculate the *prior* probabilities of class membership using a multinomial logistic regression. By default, the largest latent class is the referent group. For example, the prior probability of belonging to latent class 3 for the Weibull model is:

$$\frac{\exp(-1.0676)}{\exp(-1.0676) + \exp(0.0317) + \exp(-0.0503) + \exp(0)} = 0.1033 = 10.3\%$$

These probabilities are used to estimate the posterior probability of class assignment using Bayes Theorem, where $\hat{\pi}_{i,k} = P(c_i = k|X_i, Y_i, \hat{\theta})$ where $\hat{\theta}$ is the vector of parameters estimated in the K latent class models. The estimates for the survival model include the covariates of interest in the model (education status and sex) as well as the baseline hazard parameters such as the class-specific shape and scale parameters for the Weibull model, the class-specific proportional hazards estimates for the HARE model, and the log-splines of the hazard function for the HARE model. The longitudinal results provide estimates of education, class-specific intercepts, class-specific age estimates, and class-specific age^2 estimates of the change in normed MMSE score over the study period. The random-effects estimates for the longitudinal model provide a lower diagonal matrix of the variance-covariance matrix of the class-specific covariates. This matrix is typically labeled the **G** matrix in mixed-effects models literature.

The class-specific survival parameters are statistically significant for both the Weibull and HARE models. Certain log-spline estimates for the hazard function from the HARE model are not statistically significant. This lack of statistical significance is of minimal importance to this model as the determination of the hazard knots are based on BIC values of tensor spaces,

Table 4.2.13 Joint LCMM Model Estimates for 4-Class Solution (Weibull)

Class-Membership				
Variable	β	SE†	Wald Statistic	p-value
Intercept (Class 1)	0.0317	0.247	0.128	0.898
Intercept (Class 2)	-0.0503	0.245	-0.205	0.838
Intercept (Class 3)	-1.0676	0.456	-2.343	0.019
Survival Model				
Variable	β	SE†	Wald Statistic	p-value
$\pm\sqrt{\zeta_1}$ (Class 1)	0.101	0.001	153.224	< 0.001
$\pm\sqrt{\zeta_2}$ (Class 1)	7.075	2.132	3.319	< 0.001
$\pm\sqrt{\zeta_1}$ (Class 2)	0.105	0.001	366.569	< 0.001
$\pm\sqrt{\zeta_2}$ (Class 2)	7.030	0.598	11.763	< 0.001
$\pm\sqrt{\zeta_1}$ (Class 3)	0.101	0.001	270.157	< 0.001
$\pm\sqrt{\zeta_2}$ (Class 3)	10.219	4.021	2.542	0.011
$\pm\sqrt{\zeta_1}$ (Class 4)	0.110	0.001	222.203	< 0.001
$\pm\sqrt{\zeta_2}$ (Class 4)	5.092	0.3192	15.751	< 0.001
Education	-0.673	0.263	-2.560	0.010
Male	0.171	0.319	0.535	0.593
Longitudinal Fixed Effects				
Variable	β	SE†	Wald Statistic	p-value
Intercept (Class 1)	62.13	3.84	16.19	< 0.001
Intercept (Class 2)	65.05	5.85	11.12	< 0.001
Intercept (Class 3)	76.95	12.18	6.32	< 0.001
Intercept (Class 4)	64.28	3.78	16.70	< 0.001
Centered-Age (Class 1)	10.79	3.97	2.72	0.007
Centered-Age (Class 2)	4.20	6.54	0.64	0.520
Centered-Age (Class 3)	-17.77	14.84	-1.20	0.231
Centered-Age (Class 4)	2.96	5.78	0.51	0.609
Centered-Age^2 (Class 1)	-4.28	1.16	-3.69	< 0.001
Centered-Age^2 (Class 2)	-4.72	1.81	-2.60	0.009
Centered-Age^2 (Class 3)	1.74	3.97	0.44	0.662
Centered-Age^2 (Class 4)	-7.91	2.20	-3.59	< 0.001
Education	13.31	1.22	10.91	< 0.001

Σ	Intercept	Centered-Age	Centered-Age^2
Intercept	147.04		
Centered-Age	-44.05	34.51	
Centered-Age^2	6.96	-10.60	3.73

Score Test for Conditional Independence: 7.919 (p-value = 0.048)
† = Some SE < 0.001 rounded up to 0.001

Table 4.2.14 Joint LCMM Model Estimates for 3-Class Solution (HARE)

Class-Membership				
Variable	β	SE†	Wald Statistic	p-value
Intercept (Class 1)	0.1786	0.283	0.630	0.528
Intercept (Class 2)	0.6175	0.275	2.238	0.025
Survival Model				
Variable	β	SE†	Wald Statistic	p-value
log(spline 1)	-10.338	318.5	-0.032	0.974
log(spline 2)	-7.459	15.63	-0.477	0.633
log(spline 3)	-16.139	393.3	-0.041	0.967
log(spline 4)	-3.262	0.802	-4.068	< 0.001
log(spline 5)	-1.016	0.902	-1.126	0.260
log(spline 6)	-0.784	6.170	-0.127	0.899
log(spline 7)	3.824	0.642	5.952	< 0.001
PH estimate (Class 1)	4.075	0.748	5.445	< 0.001
PH estimate (Class 2)	-3.450	0.540	-6.413	< 0.001
Education	-0.653	0.252	-2.592	0.010
Male	0.415	0.395	1.051	0.293
Longitudinal Fixed Effects				
Variable	β	SE†	Wald Statistic	p-value
Intercept (Class 1)	63.04	3.91	16.13	< 0.001
Intercept (Class 2)	60.32	2.97	20.34	< 0.001
Intercept (Class 3)	73.91	5.13	14.40	< 0.001
Centered-Age (Class 1)	5.33	6.15	0.87	0.386
Centered-Age (Class 2)	9.79	3.33	2.94	0.003
Centered-Age (Class 3)	-6.58	6.25	-1.05	0.292
Centered-Age^2 (Class 1)	-9.07	2.43	-3.74	< 0.001
Centered-Age^2 (Class 2)	-4.38	0.97	-4.53	< 0.001
Centered-Age^2 (Class 3)	-2.12	1.85	-1.15	0.252
Education	13.54	1.20	11.25	< 0.001

Σ	Intercept	Centered-Age	Centered-Age^2
Intercept	143.55		
Centered-Age	-50.80	65.66	
Centered-Age^2	14.26	-23.67	8.69

Score Test for Conditional Independence: 5.664 (p-value = 0.129)
† = Some SE < 0.001 rounded up to 0.001

but this may provide motivation for class-specific estimates of these hazard knots. Primary school education is a statistically significant covariate in both models with relatively equal estimates ($\beta = -0.673$ in the Weibull model and $\beta = -0.653$ in the HARE model). Male sex has a larger association in the HARE model ($\beta = 0.415$) relative to the Weibull model ($\beta = 0.171$), but neither effect is statistically significant. Both estimates have roughly the same standard error. Age and age^2 were statistically significantly associated with normed MMSE scores for the high-performing class in the Weibull model (latent class 1) and its equivalent class in the HARE model (latent class 2). Further, these estimates are close in value (Weibull $\beta_{age,1} = 10.79, \beta_{age^2,1} = -4.28$; HARE $\beta_{age,2} = 9.79, \beta_{age^2,2} = -4.38$). Similar parallels can be seen comparing the poor-performing latent classes in both models (Weibull LC 4 and HARE LC 1). Age itself is not statistically significant in either model, but the quadratic effect of age is. The one distinction here is that the estimate of β for the age^2 effect in Weibull is somewhat larger (-7.91) than the estimate provided by HARE (-9.07). Education is likewise similar in both models, both in magnitude ($\beta \approx 13$ in both models) and in its statistical significance. The variance of age in the random effects of the HARE model are roughly twice as large as in the Weibull model. This could possibly result from the Weibull model having more class-specific estimates, meaning that the HARE model has more unaccounted variance in age. While the magnitudes differ, the covariance between intercept levels of normed MMSE (i.e. the baseline level) and the age variables are negative for both models. This result is unsurprising seeing the negative association between cognitive functioning and age in this sample overall.

The final comparison of note is for the score test for conditional independence (Jacqmin-Gadda, et al., 2010) in both models. One assumption for estimating the likelihood of the JLCMM is that the time-to-event and the longitudinal outcome (i.e. the time to dementia onset and normed MMSE in this sample) should be independent once the latent class is known. That is, estimates within latent classes are homogeneous and provide sufficient information for estimating both outcomes. For this hypothesis test, the assumption is that the models are jointly independent (more specifically, $H_0 : \eta = 0$, where η assesses variance not explained by the

covariates within the joint model). Rejecting this null hypothesis provides for the alternative hypothesis that there is no conditional independence/there is joint dependence between the outcomes not explained by the model. This hypothesis test is rejected for the Weibull model ($p = 0.048$), but not for the HARE model ($p = 0.129$). Therefore, the latent classes as defined by the Weibull model do not fully explain the joint outcomes, whereas they do for the HARE model. This would make inference trickier for the class-specific estimates in the Weibull model than for the HARE model.

4.3 Pupil Diameter and Time-to-Fixation on Social Regions of Face among Infants with Elevated- and Low-likelihood of Autism Spectrum Disorder

Data Description

The identification of symptoms of Autism Spectrum Disorder (ASD) typically occur towards the latter part of the first year or life or later, and diagnosis of ASD usually occurs at ages 24- to 36-months-old at earliest (Grzadzinski, et al., 2017). However, neurobiological indicators such as differences in the cortical surface area of children with an elevated likelihood (EL) of developing ASD can be detected as early as 6-months of age through the use of structural MRI (Hazlett, et al., 2017). While MRI or other imaging techniques could be used to identify prospective cases of ASD during a pre-symptomatic period of age, these technologies are expensive (Grzadzinski, et al., 2017) and difficult to implement in infants (Raschle, et al., 2012). Therefore, research has been done to find biological and behavioral markers of ASD during the pre-symptomatic period that do not rely on the use of imaging technology.

The SESAMI (Stimuli for Early Social Arousal and Motivation in Infants) paradigm was developed to assess social arousal responses in the form of pupillary, cardiac, and respiratory responses during social stimuli among EL-ASD children and a comparative group of children with a low likelihood (LL) of developing ASD. This paradigm establishes an experiment of the presentation of a series of videos to infants. A total of nine videos alternating between control and stimulus trials are presented in this series, and multiple wearable devices measure

the arousal biometrics of these infants. The stimulus trials depict videos of faces as they look forward towards the viewer and speak some social phrase to the infant. Social phrases included, "Hey baby!" or "I hope you're having a good day." These faces are placed in front of a black, neutral background. The videos are formatted into two sections–a static portion and a dynamic portion. The static portion comprises a five-second sequence of a neutral, unmoving face with no sound. The dynamic portion comprises a five-to-eight second sequence of the moving face with accompanying social audio.

Control trials operated similarly to stimulus trials with videos comprising a static portion and dynamic portion. However, to ensure that the control trials were absent of social stimuli, various alterations were made to the stimulus videos. The faces used in these videos were inverted upside-down and pixelated. Audio was filtered through a musical harmonization program that changed verbalization to lilts. (One might describe this process as "pixelating" a voice). Control trials, therefore, presented a static, pixelated image of an inverted face for five seconds followed by a dynamic pixelated face that lilted for five-to-eight seconds. Constant luminance was maintained throughout all control and stimulus trials to avoid differential pupillary response to light. The goal of this experiment was to determine what levels biometric responses of arousal were produced by social cues in the presence of other typical arousal responses.

Pupillary response data, which included gaze location on a standardized XY-coordinate system of the screen and pupil diameter for each eye, was collected with Tobii Pro Lab (v1.181) using the Tobii Pro Fusion eye tracking device (Tobii Pro AB, 2014). Data were collected at 1000 Hz. The standardized coordinate system was a media coordinate system norm with the X- and Y- coordinates ranging from 0 to 1. The origin of this screen is at the top-left position making the x-axis range from 0 to 1 in a left-to-right direction and the y-axis range from 0 to 1 in an up-to-down direction. Pupil diameter was measured in centimeters.

From these data, a total of $N = 64$ infants were included in these analyses with 17 EL infants and 47 LL infants. Down syndrome infants and their respective controls were excluded from the analyses due to low proportions of infants with non-missing pupil data. There were 32

males (11 in EL group and 21 in LL group) and 32 females (6 in EL group and 26 in LL group). The average age was 11.9 months (SD = 5.62, range = 5.9 to 25.1). For EL infants, the average age was 10.8 months (SD = 7.49, range = 5.9 to 25.1), and the average age for LL infants was 12.1 months (SD = 5.2, range = 6.23 to 23.6).

Methods

The goal of these analyses was to determine whether latent class structure could be estimated from joint outcomes of pupil diameter size and the time taken to fixate on a social region of the face. These arousal profiles may be heterogeneous during the pre-symptomatic period (Grzadzinski, et al, 2017), so a JLCMM would provide a sound method of estimating the joint model. Further, HARE would likely be preferable over a parametric model because of its looser set of assumptions for this heterogeneous population. The change in pupil size was estimated using a linear mixed-effects model, and the time-to-fixation on a social region was estimated using HARE with class-specific baseline hazards.

Since pupil diameter data were time series and infinite-dimensional, these data were transformed and down-sampled to be appropriate for a linear mixed-effects model. Over the sequence of nine video trials, some seasonality effects occurred with a number of participants where the pupil size sustained some increase over the period of the trials. In a few cases, pupil sizes sustained decrease over this period. Further, some gaps existed where the infant's eye data were not registered by Tobii, presumably due to looking away from the screen. These gaps of missing data were filled in using a last observation carried forward method. This method provided a conservative estimate of pupil size change as missing data typically occurred in troughs and occassionally at crests. After missing data were imputed, pupil diameter data were detrended using the `detrend()` function in **astsa** in R v1.4.0 (Stoffer & Poison, 2023). This detrending scaled the pupil diameters such that they had a mean of 0 and standard deviation of 0.5. Left and right pupil diameters were largely concordant at a lag of 0 milliseconds (CCF = 0.611), so the left pupil diameter was arbitrarily chosen as the longitudinal outcome. These data were then downsampled to every 2500 milliseconds (ms) to accommodate the mixed-effects model.

Time-to-fixation was defined as the time in ms taken to look at any social region or regions for a total of 2000 ms (2 seconds). Social regions included the eyes (x-range from 0.35 to 0.65, y-range from 0.4 to 0.47), nose (x-range from 0.45 to 0.60, y-range from 0.47 to 0.55), or mouth (x-range from 0.40 to 0.62, y-range from 0.62 to 0.75). The 2000 ms time threshold was determined using a few criteria. Analyses for the SESAMI validation paper (paper forthcoming) indicated that mean fixation time for any social region of interest was approximately 4000 ms. However, any one social region had, on average, a total fixation time of 1500 ms. Therefore 2000 ms offered a time that likely indicated a participant had fixated on multiple social regions. Further, as shown in Figures 4.3.1 and 4.3.2, a 2000 ms fixation period offered an excellent amount of variation compared to other amounts of time. A 1000 ms total was too low a threshold for the participants and 4000 ms onward had too few observed outcomes. This latter issue would be especially difficult in trials where participants began to fixate on a social region towards the latter half of the dynamic section of the video. Some dynamic portions of the video only lasted 5 seconds.

Only stimulus trials were used in the analyses, providing nine trials for each participant. This decision was made because there was too little variability in the time-to-fixation in the control tirals. In many of these trials, several participants did not fixate on any social region. On average, fixation on any social region of the "face" during the control trial lasted slightly over one second. This fixation time was less than 500 milliseconds on average for any individual social region.

The longitudinal model for this analysis was

$$y_i(t) = \beta_{0k} + \beta_{1k}x_{i1} + \beta_{2k}x_{i1}^2 + \beta_{3k}x_{i1}^3 + \beta_4 x_{i2} + \beta_5 x_{i3} + \beta_6 x_{i4} + \beta_7 x_{i5} + \omega_i(t) + \epsilon_i(t),$$

where $y_i(t)$ is the detrended left pupil diameter for participant i at time t, $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \beta_{2k}, \beta_{3k})$ is the vector of class-specific fixed-effects estimates for latent class k, $\boldsymbol{\beta} = (\beta_4, \beta_5, \beta_6, \beta_7)$ is the vector of marginal fixed-effects estimates, $\omega_i(t)$ is afirst-order autoregressive stationary process

such that $\text{cov}(\omega_i(t), \omega_i(s)) = \sigma^2_\omega \exp\{(-\rho|t-s|)\}$, and $\epsilon_i(t)$ is the random error for participant i at time t. Due to the periodic nature of the pupil diameter change, $\omega_i(t)$ accounted for the within-subject variance in lieu of random effects. The independent variables $x_{i2}, x_{i3}, x_{i4}, x_{i5}$ were estimated marginally rather than by each class k. Variable x_{i2} was age of the participant in months, x_{i3} was sex with female as the referent category, x_{i4} was the order of the video (ordinal ranging 1 to 9), and x_{i5} was the likelihood group with LL as the referent category. The independent variable x_{i1} represented the cumulative time in the trial measured in milliseconds. To approximate the periodic nature of the pupil diameter over this time period, a cubic orthogonal polynomial for x_{i1} was estimated using the `poly()` function in base R (R Core Team, 2021). The survival model was estimated as

$$\lambda_i(t) = \lambda_{0k}(t) \exp\{\theta_{1k}x_{i1} + \theta_{2k}(x_{i1} - 19.6)_+ + \theta_{3k}(x_{i1} - 23.13)_+ + \theta_{4k}x_{i5}\}.$$

Both x_{i1} and x_{i5} are the same as in the longitudinal model. The parameter θ_{4k} is a class-specific estimate of the effect of likelihood group on time-to-fixation. The age variable x_{i1} was split into a linear B-spline with class-specific estimates for each knot. θ_{1k} is the class-specific estimate for age, θ_{2k} is the class-specific estimate for age after the first knot of 19.6 months, and θ_{3k} is the class-specific estimate for age after the second knot of 23.13 months. The survival model was determined by using HARE and including age, sex, likelihood group, and video order as potential independent variables. HARE was performed using the `hare()` function in the **polspline** package in R v4.1.0 (Kooperberg, 2020). HARE determined age with knots at 19.6 months and 23.13 months were influential to the survival model as was likelihood group. Further, HARE determined knots in time should be indicated at 3482 ms and 6958 ms. Latent class membership was modeled as intercept-only.

Models with 1 - 3 latent classes were tested. For the two- and three-class models, a second model was tested where likelihood group was included in the latent class membership model. The JLCMMs were estimated using the `Jointlcmm()` function from the **lcmm** package

in R v4.1.0 (Proust-Lima, et al., 2017). All models used the `hazard = ''Specific''` option and `hazardnodes = c(3482, 6958)` option to correspond to the knots in time. Parameters were estimated using the `gridsearch()` function with the one-class solution as the initialization model, 30 repetitions used for each iteration, and a maximum of 15 iterations specified. During the estimation process, convergence issues occurred with the Levenberg-Marquardt algorithm such that the second derivative of the Hessian threshold could not be met. This problem often occurs when the solutions to estimates are on the boundary of support (Proust-Lima, et al., 2017). The specific convergence issue indicated that the variance estimate for some coefficients was 0 (i.e. on the boundary of support for variance estimates). Checking the `V` matrix from the `Jointlcmm()` output verified this problem for some variables estimating the class-specific baseline hazard functions over time. For these variables, fixing their position with the `posfix` option prompted convergence of the Levenberg-Marquardt algorithm. These variables are noted in Table 4.3.6 with an asterisk.

JLCMM predicted outcomes for $y_k(t)$, $\lambda_k(t)$, and $S(t)_k$ were made after the optimal number of latent classes was determined. To illustrate the advantages of a JLCMM, a Cox proportional hazards model was estimated using the detrended pupil diameter as a time-varying covariate. Since pupil diameter was time-varying, using an interaction with the latent classes as a separate covariate could have violated the proportionality assumption. As such, the data were decomposed into separate datasets for each separate latent class. Then a Cox proportional hazards model was run on the separate data sets. The Cox model was estimated using `coxph()` from the **survival** package in R (Therneau, 2021).

Results

Figure 4.3.1 shows the distribution of times-to-event using different durations of fixation on social regions for "Heather Good Day" video. Figure 4.3.2 shows these distributions for "Jaclyn Smile" videos. The primary variability existed with a 1000 and 2000 ms threshold. Some variability existed for 3000 ms with "Jaclyn Smile" and a few other videos, but many looked more similar to the "Heather Good Day" distribution. Since 1000 ms was considered too low

a threshold, and because of the results from the validation paper, 2000 ms was selected as the duration needed to be considered fixated on a social region.

Table 4.3.1 provides a summary of the fit statistics for the five models tested: a one-class solution, two two-class solutions, and two three-class solutions. The models labeled with "CM" included an additional parameter in the model that estimated the multinomial logistic regression latent class membership model with an independent variable for likelihood group. Minimal BIC was used to determine the model with the best fit. The optimally fiting model had three classes with 26 members in LC 1 (40.6%), 18 members in LC 2 (28.1%), and 20 members in LC 3 (31.3%). A few of these models have BIC values close together indicating that any of these models may be suitable. Part of the decision to select the three-class model, aside from its BIC, is the fact that it improved more compared to the CM model than the two-class solution did. That is, the two-class and two-class CM models were relatively similar in fit, but the three-class model was much better than the three-class model with CM. Table 4.3.2 provides a summary of the mean posterior probability of assignment for each latent class. Discrimination between LC 2 and LC 3 was reasonable, with the mean probabilities of assignment into these classes being greater than 0.7. The mean probability of assignment to LC 3 among those who were assigned to LC 1 was about 22%, indicating some similarities in the joint outcomes of these two latent classes.

Table 4.3.1 Joint Latent Class Mixed Model Fit Summaries

Model	K	LL	npm	BIC	% Class 1	% Class 2	% Class 3
One Class	1	-1304.1	21	2750.6	100%		
Two Class	2	-1293.5	36	2736.8	35.9%	64.1%	
Two Class with CM[1]	2	-1291.5	37	2737.0	26.6%	73.4%	
Three Class	3	-1283.0	42[2]	2732.4	40.6%	28.1%	31.3%
Three Class with CM	3	-1293.3	53	2807.1	25.0%	53.1%	21.9%

1 = CM refers to "Class Membership" variable added (Likelihood Group)
2 = Some parameters fixed to allow for convergence

Table 4.3.3 provides further assessment of the discriminatory power of these three latent class assignments. The proportion of posterior probabilities above three thresholds, 0.7, 0.8,

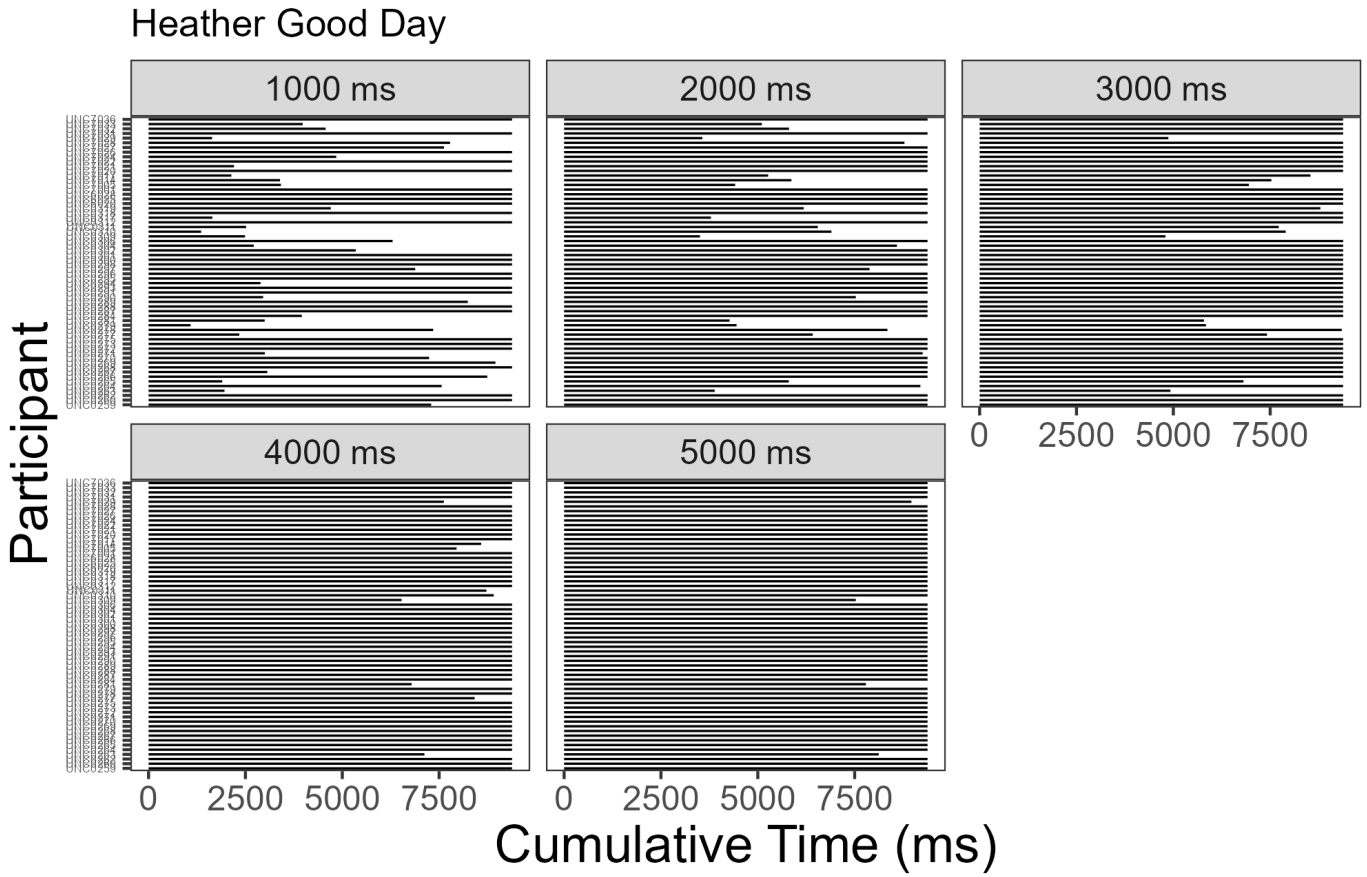

Figure 4.3.1 Bar charts of Time-to-Event for Different Durations of Fixation on Social Regions for “Heather Good Day” Stimulus

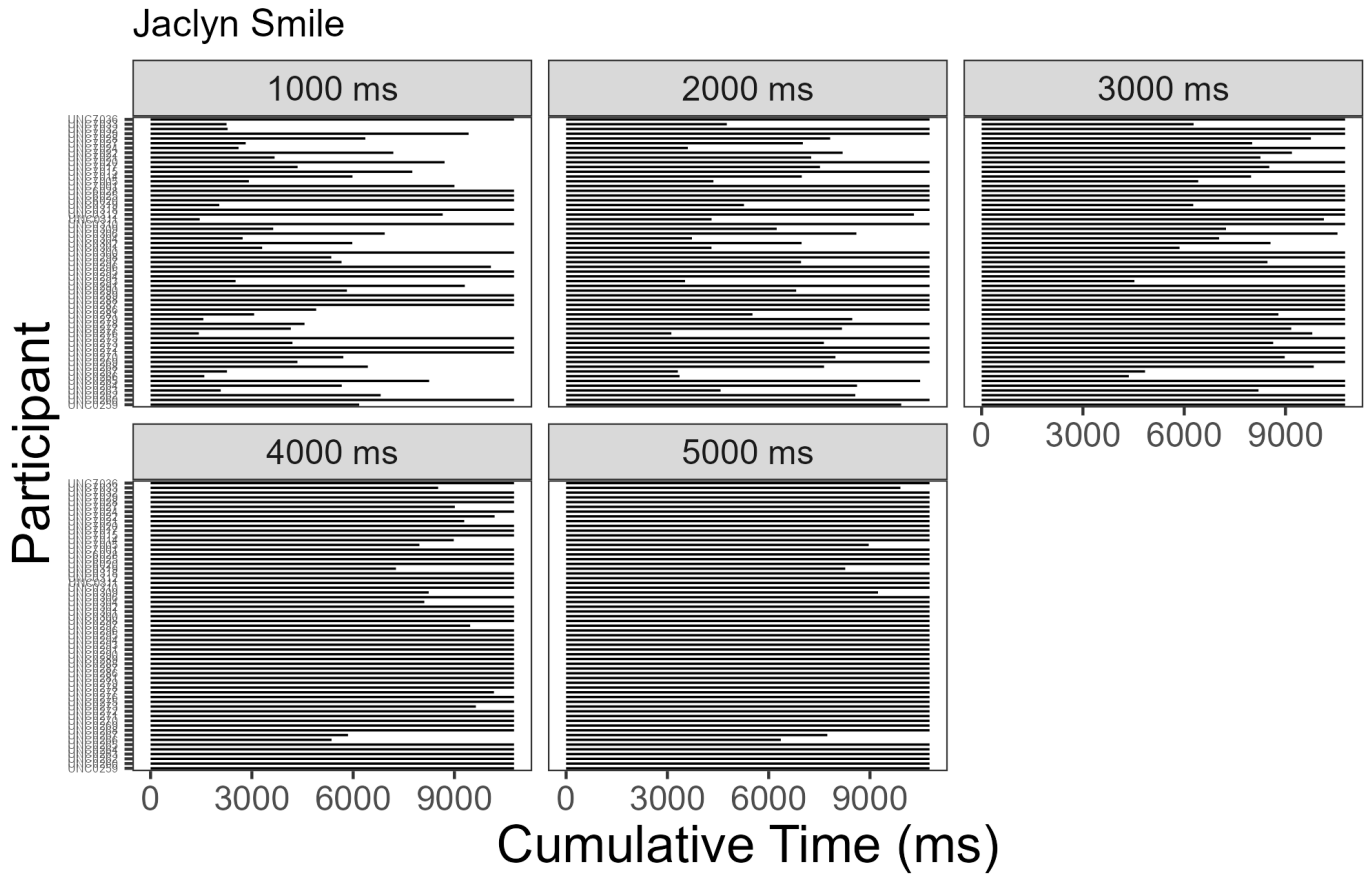

Figure 4.3.2 Bar charts of Time-to-Event for Different Durations of Fixation on Social Regions for “Jaclyn Smile” Stimulus

and 0.9, are provided in the table. LC 2 has great discriminatory power, with at least 60% of LC 2's posterior probabilities of assignment into class 2 being above 0.8. Similarly, LC 3 has great discriminatory power with 45% of its posterior probabilities of assignment into class 3 being above 0.9. Contrarily, LC 1 has lower power in its discrimination with only 46% of its posterior probabilities being above 0.7 and slightly more than 10% above 0.9 probability. These results for LC 1 indicate that some characteristics are shared between LC 1 members with LC 2 and LC 3.

Table 4.3.2 Mean of Posterior Probabilities for Three-Class HARE Model

	$P(c_i = 1)$	$P(c_i = 2)$	$P(c_i = 3)$
1	0.685	0.098	0.218
2	0.114	0.821	0.065
3	0.137	0.082	0.781

Table 4.3.3 Proportion of Posterior Probabilities above Various Thresholds

	k = 1	k = 2	k = 3
$P(c_i = k) \geq 0.7$	46.2%	77.8%	65.0%
$P(c_i = k) \geq 0.8$	23.1%	61.1%	50.0%
$P(c_i = k) \geq 0.9$	11.5%	38.9%	45.0%

Figures 4.3.3, 4.3.4, and 4.3.5 show the predicted longitudinal outcomes, the predicted hazard functions, and the predicted survival functions for each latent class, respectively. LC 1 and LC 2 had periodic trajectories of pupil diameter chage with LC 1 having a more dynamic change than LC 2. LC 3 had a muted pupillary response over time beginning with higher levels than either other LC and having the lowest overall pupil size by the end of the trial. Figure 4.3.6 illustrates a more thorough examination of pupil size over the trial time, with observations from each trial aggregated on the figure. Note that the points are jittered for decipherability.

LC 2 had a higher hazard of social fixation over the course of the trials than either other group. This hazard was mostly increasing, leading to a low probability of this LC not having fixated on social regions by the end of each trial. LC 3 had a hazard of fixating on social regions that peaked around 7000 ms leading to a roughly 50% probability of going through any trial

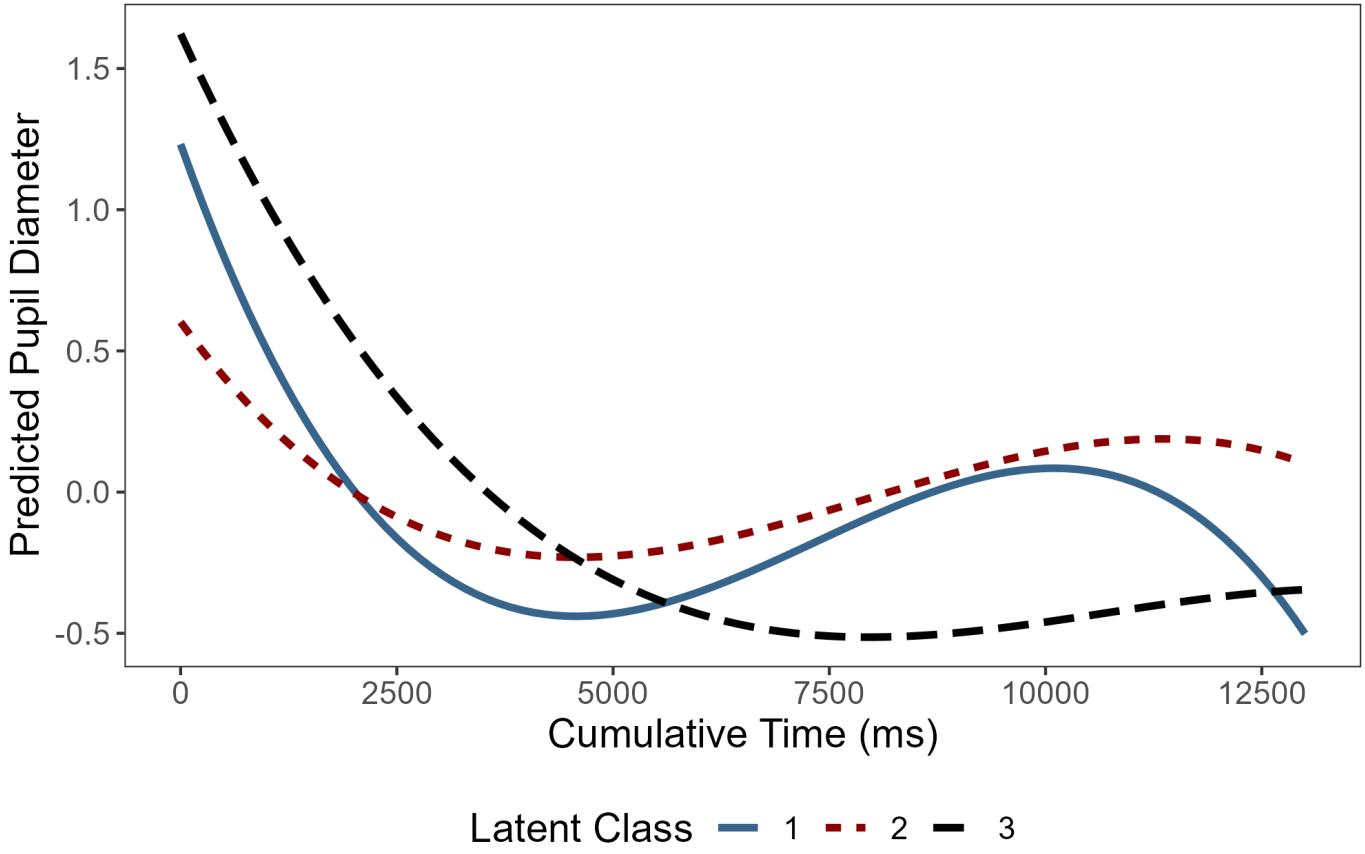

Figure 4.3.3 JLCMM Model-Predicted Detrended Pupil Size

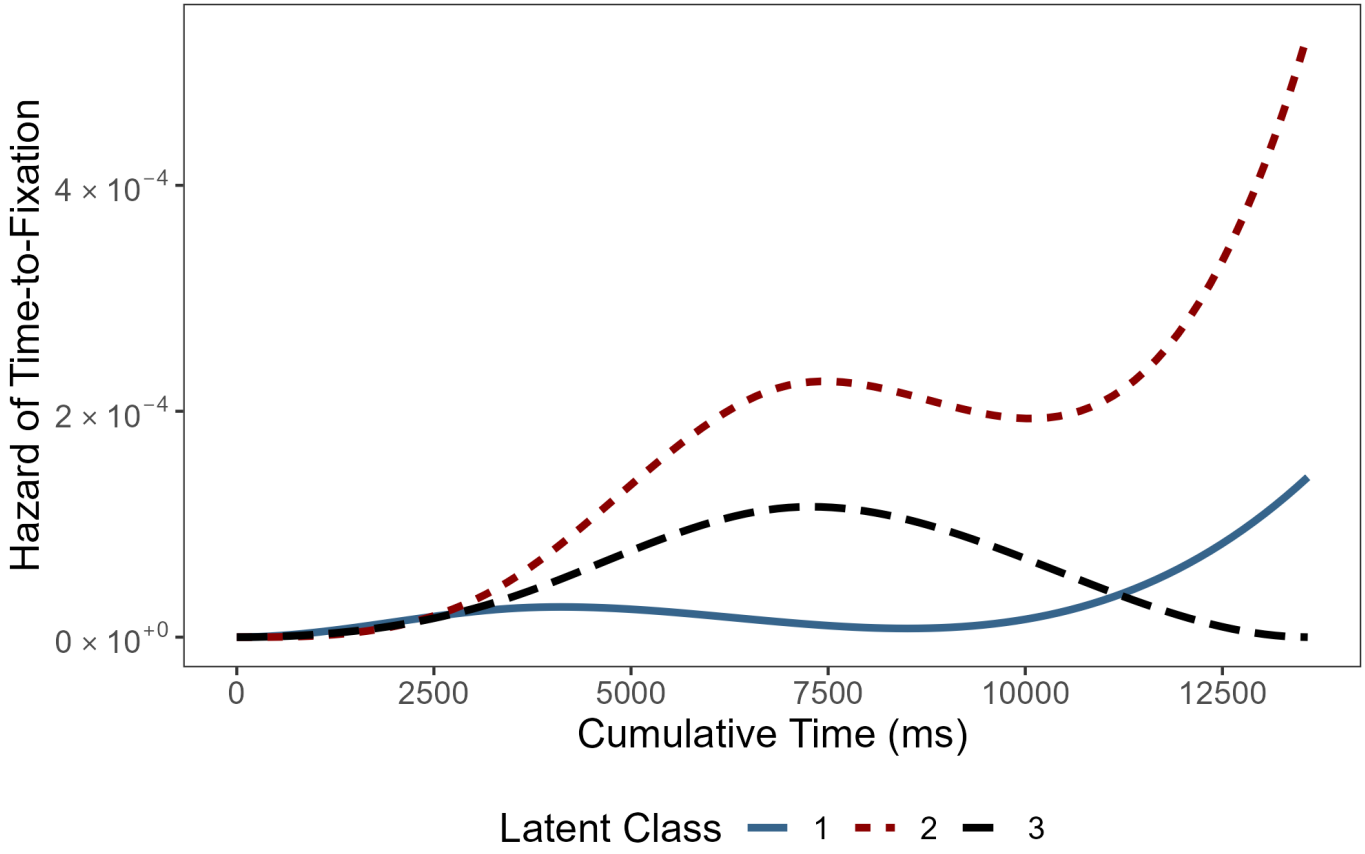

Figure 4.3.4 JLCMM Model-Predicted Hazard of Fixating on Social Region for 2000 ms

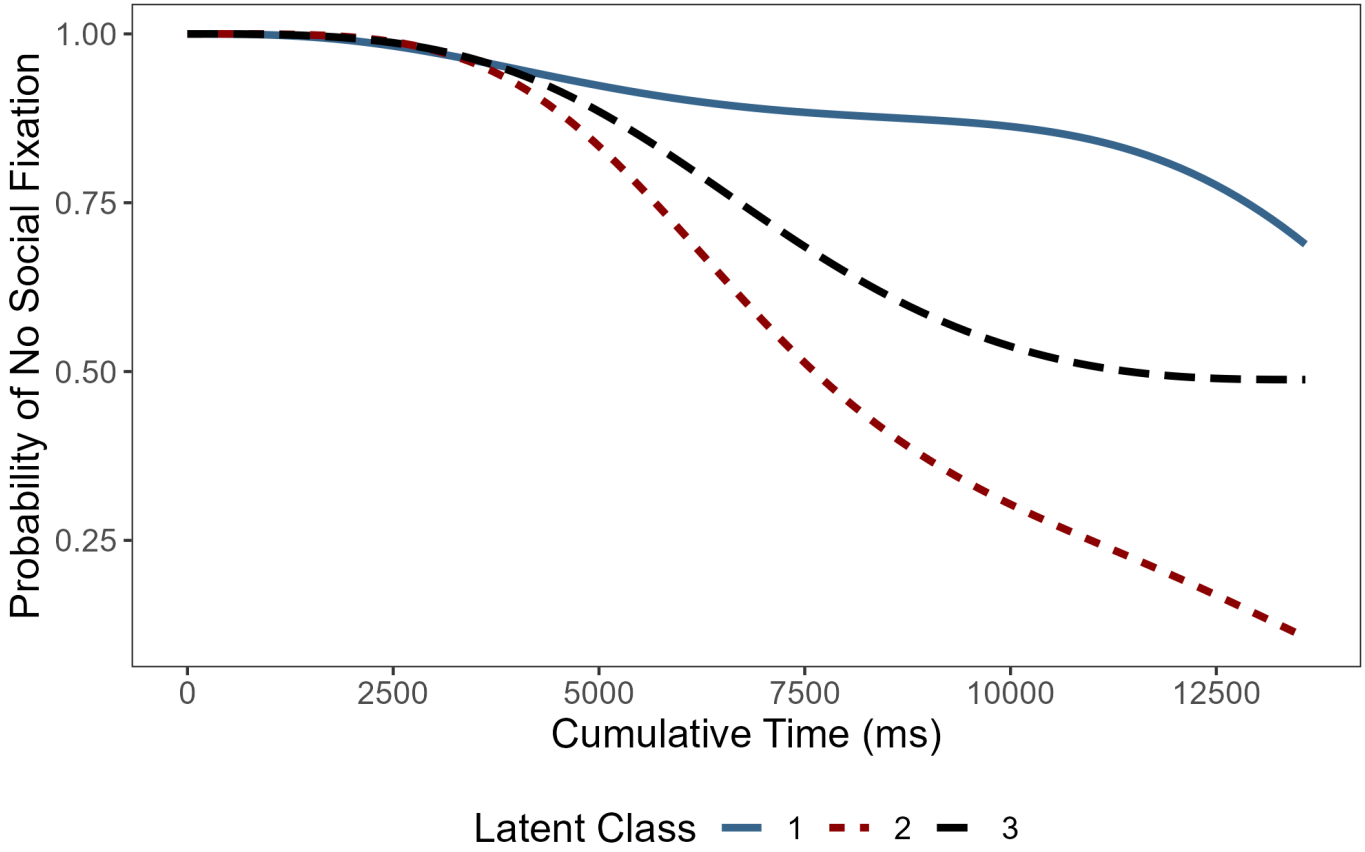

Figure 4.3.5 JLCMM Model-Predicted Probability of Avoiding Facial Social Region Fixation

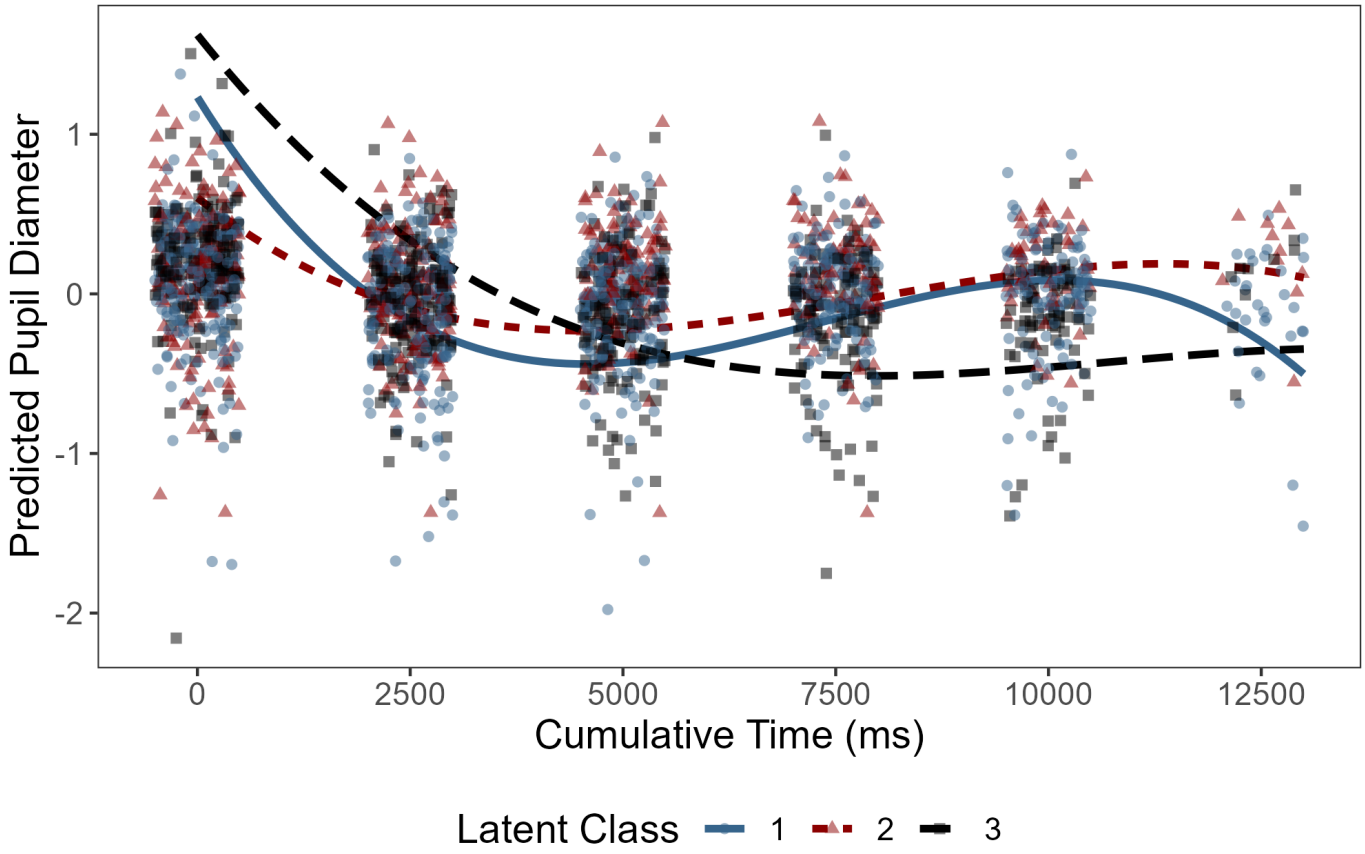

Figure 4.3.6 JLCMM Model-Predicted Detrended Pupil Size for Each Latent Class with Observed Pupil Sizes overlayed. Points are jittered.

without having fixated on social regions. LC 1 had a very low hazard of every fixating on social regions with few participants reaching the 2000 ms threshold. Table 4.3.4 shows the contingency table of event versus censoring over the course of the nine trials. As indicated by the hazard and survival plots, LC 1 had fewer observed social fixations than either LC 2 or LC 3 relative to class size with 59% of the total observations being censoring. For comparison, 33% of LC 2 had censored observations, and 47% of LC 3 had censored observations. Assuming trials are independent, which is unlikely, the χ^2 test indicates that latent class membership is not independent of censoring ($\chi^2 = 23.9, df = 2, p < 0.001$). This association between latent class membership and censoring was more apparent by the final trial (Table 4.3.5), where the χ^2 test indicates an association between latent class membership and censoring ($\chi^2 = 18.6, df = 2, p < 0.001$) with LC 1 being much more likely to have censored data compared to either LC 2 (OR = 14.7) or LC 3 (OR = 9.8).

Table 4.3.4 Observed Event For Social Fixation over Nine Trials

K	Observed	Censored	
1	90	129	
2	104	52	
3	89	79	
			χ^2 = 23.9

Table 4.3.5 Observed Event For Social Fixation over Final Trial

K	Observed	Censored	
1	5	21	
2	14	4	
3	14	6	
			χ^2 = 18.6

The predicted outcomes provide some details about the characteristics of the latent classes. LC 1 largely comprises kids who do not fixate on social regions. The periodic nature of their pupil diameter change may be a natural pupillary cycle or a remnant of arousal due to auditory cues. LC 1 appeared to lose attention over the course of the trials with fewer members fixating

on social regions by trial 9. LC 2 had the highest hazard of fixating on social regions and had a sustained positive pupillary response over the course of the trial. The probability of not fixating on social regions by the end of the trial was small ($S(t)_2 = 0.106$)) for this latent class, and this behavior persisted throughout the trials. LC 3 had a muted pupillary response to these trials, but also had a non-zero probability of fixating on social regions. For LC 3, the probability of not fixating on social regions by the end of a trial was roughly 50% ($S(t)_3 = 0.488$). However, their attention was similar to LC 2 in that fixation was more observed by the final trial than censoring.

Table 4.3.6 provides β coefficient estimates for the linear mixed-effects model and θ coefficient estimates for the log-hazard model results from the JLCMM. Some estimates, provided with an asterisk, had to be fixed in order to ensure convergence of the Levenberg-Marquardt algorithm to the maximum likelihood estimate. The variance of these estimates were exactly 0, indicating the conjectured boundary estimation problem was correct. The zero variance of these spline estimates indicated potential limitations of the HARE method indentifying knots in splines where not all log-hazard functions between classes were similar. These limitations are discussed in Chapter 5. Otherwise, the log-hazard estimates of these splines follow what was shown in Figures 4.3.4 and 4.3.5. ASD likelihood was not statistically significantly different in estimating the log-hazard with the exception of LC 2, where those with EL of ASD diagnosis had a much higher log-hazard estimate of social region fixation than their LL counterparts ($\theta_{42} = 4.31, SE = 1.46, p = 0.003$). Age in months also had some different effects between the latent classes. Older participants had the highest log-hazard of social region fixation except within the interval between 19.6 and 23.13 months, which saw no real change. Note the differences in effects from the knots in age for different latent classes. LC 2 shows a strong, statistically significant effect for the knot at 19.6 months. LC 3 shows a strong, but not statistically significant, effect for the knot at 23.13 months. These knots were determined by HARE in a data-driven fashion and would not have been noticed unless specifically tested in either a parametric model or a non-parametric model without *a priori* determination of these knots.

Table 4.3.6 Estimates of Coefficients for the Survival Model and Linear Mixed-Effects Model of the JLCMM

Survival Model				
Parameter	θ	SE	t-value	p-value
(spline 1)$^{1/2}$ LC 1*	-0.00004	X	X	X
(spline 2)$^{1/2}$ LC 1*	-0.00002	X	X	X
(spline 3)$^{1/2}$ LC 1	0.633	0.218	2.90	0.004
(spline 4)$^{1/2}$ LC 1*	0.0001	X	X	X
(spline 5)$^{1/2}$ LC 1*	0.0011	X	X	X
(spline 6)$^{1/2}$ LC 1	0.0001	0.212	0.00	0.999
(spline 1)$^{1/2}$ LC 2*	0.0003	X	X	X
(spline 2)$^{1/2}$ LC 2*	0.0003	X	X	X
(spline 3)$^{1/2}$ LC 2*	0.0235	X	X	X
(spline 4)$^{1/2}$ LC 2*	1.1578	X	X	X
(spline 5)$^{1/2}$ LC 2	-0.7906	1.759	-0.449	0.653
(spline 6)$^{1/2}$ LC 2	4.001	2.395	1.670	0.095
(spline 1)$^{1/2}$ LC 3*	-0.0001	X	X	X
(spline 2)$^{1/2}$ LC 3	0.0003	0.024	0.00	0.999
(spline 3)$^{1/2}$ LC 3	0.0001	0.052	0.00	0.999
(spline 4)$^{1/2}$ LC 3	0.1539	0.269	0.573	0.567
(spline 5)$^{1/2}$ LC 3	0.7089	0.284	2.498	0.013
(spline 6)$^{1/2}$ LC 3*	0.003	X	X	X
Age LC 1	0.917	1.132	0.810	0.419
Age LC 2	4.245	1.721	2.466	0.014
Age LC 3	7.814	4.334	1.803	0.071
(Age - 19.60)$_+$ LC 1	1.720	1.985	0.866	0.386
(Age - 19.60)$_+$ LC 2	5.726	1.693	3.381	< 0.001
(Age - 19.60)$_+$ LC 3	-0.944	15.086	-0.00	0.999
(Age - 23.13)$_+$ LC 1	-1.745	99.521	-0.00	0.999
(Age - 23.13)$_+$ LC 2	-0.962	2.004	-0.480	0.631
(Age - 23.13)$_+$ LC 3	6.024	4.785	1.259	0.208
Elevated Likelihood ASD LC 1	0.052	0.855	0.061	0.952
Elevated Likelihood ASD LC 2	4.308	1.458	2.954	0.003
Elevated Likelihood ASD LC 3	0.807	1.420	0.568	0.570
Mixed-Effects Model				
Parameter	β	SE	t-value	p-value
Intercept LC 1	-0.064	0.028	-2.163	0.031
Intercept LC 2	0.021	0.0374	0.561	0.576
Intercept LC 3	-0.099	0.034	-2.936	0.003

Parameter (cont'd)	β	SE	t-value	p-value
Time LC 1	-2.189	0.607	-3.607	< 0.001
Time LC 2	0.528	0.779	0.677	0.498
Time LC 3	-4.793	1.045	-4.587	< 0.001
Time^2 LC 1	0.777	0.597	1.303	0.193
Time^2 LC 2	2.395	0.794	3.015	< 0.001
Time^2 LC 3	3.374	1.164	2.898	< 0.001
Time^3 LC 1	-1.675	0.574	-2.921	0.003
Time^3 LC 2	-1.457	0.763	-1.912	0.056
Time^3 LC 3	-0.859	1.176	-0.730	0.465
Age	0.008	0.002	0.491	0.624
Male	0.016	0.018	0.916	0.359
Video Order	0.007	0.003	2.410	0.016
Elevated Likelihood ASD	-0.0104	0.020	0.510	0.609
ω^*	0.0009	X	X	X
$\sigma^\dagger_\omega$	0.000	0.025	–	–
Residual $\sigma^\dagger_\omega$	0.3683	0.006	–	–

$*$ = Estimate fixed to ensure convergence of Hessian matrix
† = Measurements of uncertainty not calculated for these values

The linear mixed-effects model also indicates some differences between the latent classes when estimating pupil diameter change. LC 1 and LC 2 show some similarities in pupil diameter change with time with the exception of LC 2 having a more profound and sustaining increase over time (indicated by positive linear and quadratic effects of time). The trajectory seen for LC 3 appears to largely have linear and quadratic effects, but not cubic effects ($\beta_{33} = -0.859$, SE = 1.176), which corresponds to its shape not resembling a periodic function. None of age, sex, or ASD likelihood had any association with pupil diameter once latent class and time were accounted for, but video order had a slight positive association ($\beta_4 = 0.007$, SE = 0.003, p-value = 0.016). Since this value was small relative to the residual standard error, even if change is entirely linear over the course of the nine trials ($9 \times 0.007 = 0.063$), this effect may be either spurious or an artifact of the detrending process.

Table 4.3.7 shows the distribution of latent classes across likelihood groups, sex, and age groups. A few characteristics stand out that invite discussion. While EL ASD participants compose roughly 30% of LC 1 and LC 3, they compose only 17% of LC 2. LC 2 was defined

by its sustained, larger pupillary response pattern with high probability of social fixation. The dominance of LL ASD kids in this latent class provides some evidence that EL ASD kids have atypical pupilary response in conjunction with aversive social fixation. The second characteristic to note is the inverted ratio of male to female participants in LC 1 compared to LC 2 or LC 3. Boys are more likely to be members of LC 1 than girls are, and boys are more likely to be assigned to LC 1 than they are to LC 2 or LC 3. Girls are equally likely to be assigned to any LC. Furthermore, of the 8 EL ASD kids in LC 1, 6 are male and 2 are female. In fact, the female participants maintain an almost identical ratio of EL-to-LL ASD groups within each class (Table 4.3.8). Lastly, members of LC 1 are approximately 2 months younger on average than for either LC 2 or LC 3. LC 1 also has the largest range of ages, but this is not necessarily unusual since it is also the largest of the latent classes.

Table 4.3.7 Characteristics of the Latent Classes Determined by the JLCMM

Characteristic	LC 1 (n = 26)	LC 2 (n = 18)	LC 3 (n = 20)
EL ASD	8	3	6
LL ASD	18	15	14
Male	15	8	9
Female	11	10	11
Mean Age (months)	10.6	12.5	12.9
SE Age (months)	4.8	6.4	5.7
Range Age (months)	5.9 - 25.1	6.7 - 24.8	6.2 - 23.6

Table 4.3.8 ASD Likelihood Group Characteristics of the Latent Classes Determined by the JLCMM

Characteristic	LC 1 (n = 26)	LC 2 (n = 18)	LC 3 (n = 20)
Male			
EL ASD	6	1	4
LL ASD	9	7	5
Female			
EL ASD	2	2	2
LL ASD	9	8	9

A Cox proportional hazards model (Cox, 1972) was used to estimate the hazard of social fixation for 2000 ms while controlling for change in pupil diameter as a time-varying covariate. Figure 4.3.7 shows the survival curve estimates of not fixating on a social region for the three latent classes. Survival estimates are underestimated for LC 1 and LC 3, which show a lower survival probability of not fixating on a social region than found in the JLCMM. The opposite is true for LC 2, which overestimated the survival probability. More members of LC 2 had fixated on a social region by the end of the stimulus trial than is shown in 4.3.7.B. The corresponding Cox PH regression coefficients (Table 4.3.9) demonstrate the root of these issues: the effect of the pupil diameter tends towards 0 for all three latent classes. This corroborates the work of Prentice (1982) and justifies the use of joint models in general (De Gruttola, et al., 1991; Pawitan & Self, 1993). A further deficit for this Cox proportional hazards model is that the knot at age 19.6 months could not be estimated for LC 1, likely missing an important transition in the hazard function due to age.

Discussion

The analyses indicated that three latent classes may exist in the sample, and some examination of the characteristics have supported that. The three latent classes distinguish themselves from each other based on joint trajectories of pupil diameter change and fixation on social regions during a social arousal task. These latent classes can briefly be defined as a class that has early fixation and high pupil response, a class that has delayed fixation and a muted pupil response, and a class that has practically no social fixation and average pupil size change. The latter class comprises several boys who have elevated likelihood of developing ASD. The latent class with early fixation and high pupil response comprises children with low likelihood of developing ASD with an even ratio of boys and girls. The hazard model results indicated that times of interest occurred near the hazard knots for certain latent classes, and the knots in age also demonstrated to be of interest in distinguishing the latent classes and describing the association between age and social fixation. Class-specific estimates of the time variables were distinct in their predicted trajectories of pupil diameter change over the course of a stimulus trial

Table 4.3.9 Cox Proportional Hazards Models of Pupil Diameter on Social Fixation for the Three Latent Classes

Latent Class 1				
Parameter	β	SE	t-value	p-value
Pupil Diameter (detrended)	-0.0494	0.0234	-2.11	0.035
Age	1.635	0.480	3.40	0.0007
$(\text{Age - } 19.60)_+$	–	–	–	–
$(\text{Age - } 23.13)_+$	-0.017	2604.0	-0.00	0.994
Elevated Likelihood ASD	0.028	0.270	1.05	0.294
Latent Class 2				
Parameter	β	SE	t-value	p-value
Pupil Diameter (detrended)	0.037	0.229	0.161	0.872
Age	1.314	0.369	3.56	0.0004
$(\text{Age - } 19.60)_+$	0.960	0.328	2.92	0.003
$(\text{Age - } 23.13)_+$	0.624	0.429	1.45	0.146
Elevated Likelihood ASD	0.710	0.319	2.22	0.026
Latent Class 3				
Parameter	β	SE	t-value	p-value
Pupil Diameter (detrended)	-0.0632	0.0245	-2.58	0.009
Age	1.42	0.044	3.189	0.001
$(\text{Age - } 19.60)_+$	1.092	0.055	1.972	0.049
$(\text{Age - } 23.13)_+$	-16.56	217.4	-0.008	0.994
Elevated Likelihood ASD	0.033	0.326	1.02	0.307

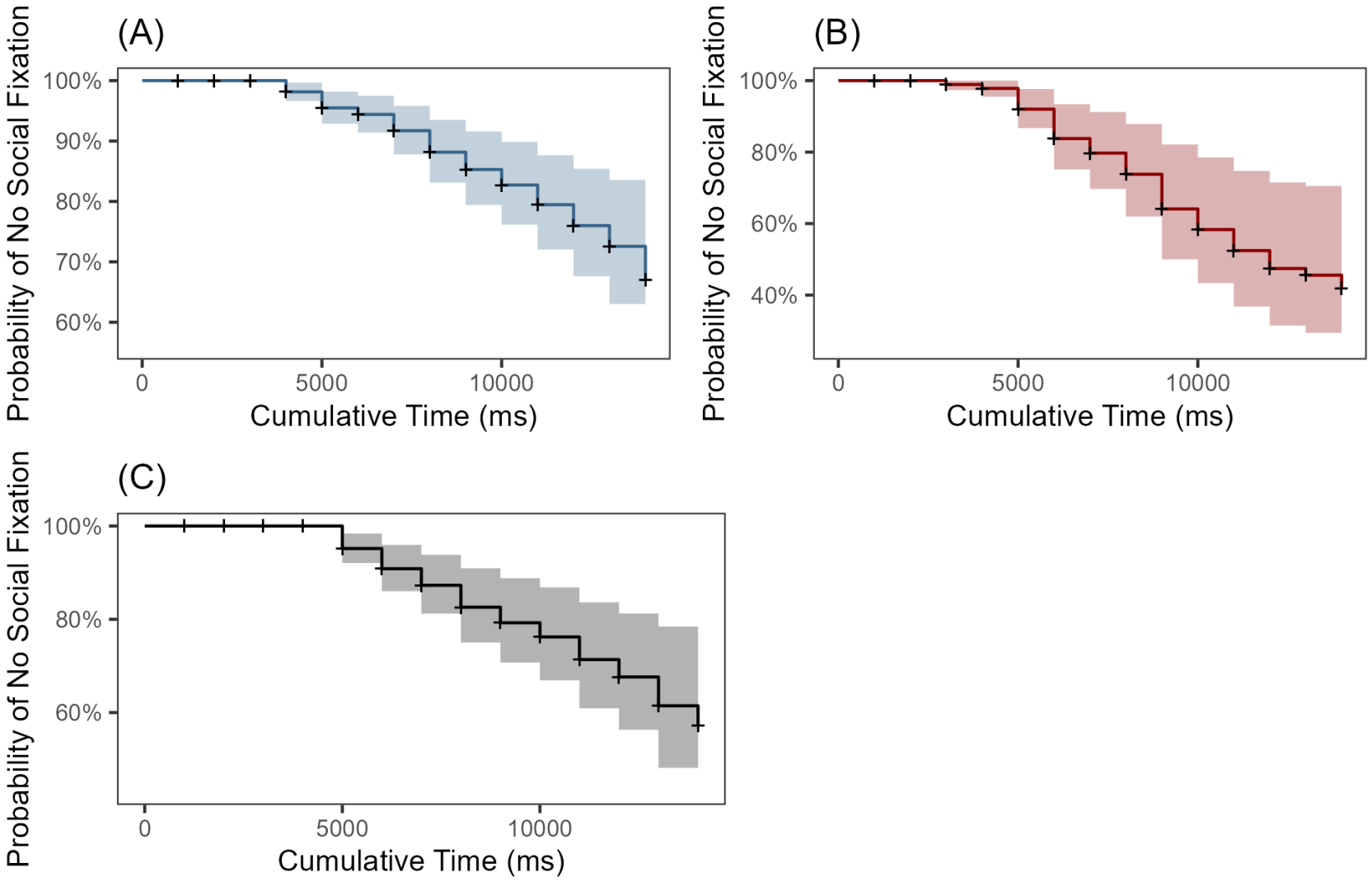

Figure 4.3.7 Predicted Survival Curves from Cox PH Models for (A) Latent Class 1, (B) Latent Class 2, and (C) Latent Class 3

for the three classes. HARE provided the class-specific estimates and appropriate knots for the hazard model without using any assumptions about the data. The proportionality assumption was relaxed, which likely aided in the estimation process as these three latent classes did not appear proportional in hazard trajectories. To the contrary, a Cox proportional hazards model using pupil diameter change as a time-dependent covariate incorrectly estimated the survival probabilities of no social fixation in a trial. These results make sense in the presence of previous literature on joint models in general. Further, this model required knowledge of latent class membership and knot location for age *a priori*, which is an optimistic assumption.

One deficit of HARE for this process was the difficulty in converging on a global maximum likelihood estimate of the vector of parameters for the JLCMM. Some spline estimates had to be fixed on estimates in order to estimate the Hessian matrix solely because the Levenberg-Marquardt algorithm could not estimate along the boundary of variance estimation. While some trial-and-error resolved this issue, a step HARE implementation should take in the future to overcome this limitation is to find a way to weight or determine knots within latent classes such that areas of sparsity can influence knot location (i.e., by avoiding these areas). Certain nonparametric methods for determining medioids, such as the k-nearest neighbor algorithm (Lloyd, 1982), could motivate such a way to choose knot locations where a large enough neighborhood is required for any latent class in order to be considered for inclusion in the hazard model.

The assessment of arousal could also be improved by incorporating other biometric measures into the longitudinal model. Pupil diameter change was used to estimate social arousal because it was the primary outcome of interest for diagnostic purposes. However, heart rate variability and respiratory rate could also lend themselves to the task of constructing an arousal metric. Such an outcome could be estimated using a state-space model (Kalman, 1960) or principal components regression (Massy, 1965).

4.4 The Cascading Effects of Reaching Major Motor Milestones on Social Behavior for Infants with Autism Spectrum Disorder

Data Description

Infants reach several developmental milestones known as "watershed moments" (Bradshaw, Schwichtenberg, & Iverson, 2022) that alter a child's interaction and behavior with his or her environment. These watershed moments include sitting upright, crawling, exhibiting selective attention, babbling, speaking, and walking, among others. These moments are not only important because of their indication of latent developmental advancements, but also because the occurrence of these watershed moments influence the occurrence of others. For example, the advancement to upright sitting leads to several new behavioral interactions, such as having more engaged shared play with a caregiver, having a larger field of vision for selective attention, and having a richer experience of interaction with objects (Bradshaw, Schwichtenberg, & Iverson, 2022). These new behaviors cascade into new watershed moments. Engaged shared play with a caregiver motivates the development of crawling, which leads to new behaviors, which in turn lead to new watershed moments and so forth. The significance of these wathershed moments and their cascading effects to developmental psychologists, particularly those who study neurodevelopmental disorders, involves understanding how a delay in these moments may compound the delay of other moments. Among children with Autism Spectrum Disorder (ASD), the heterogeneous profiles may be influenced by these delayed developmental trajectories. For example, an infant with ASD whose profile exhibits primarily restricted and repetitive behaviors may have a different developmental trajectory of motor watershed moments from another infant with ASD whose profile exhibits primarily joint attention issues.

Data were acquired from an NIH-funded Autism Centers of Excellence (ACE) network study commonly referred to as the "Infant Brain Imaging Study" (IBIS) (Hazlett, et al., 2017). This network includes four clinical data sites from across the United States (UNC-Chapel Hill, University of Washington, Washington University in St. Louis, and the Children's Hospital of Philadelphia), two image processing sites (UNC-Chapel Hill and University of Utah), and a data coordinating center (McGill University). Infants with an elevated-likelihood (EL) and

low-likelihood (LL) of developing ASD were enrolled at 6-months of age and followed through school age. Data collected at each visit include functional and structural MRI scans, behavioral assessments, and biological assays. More information about the IBIS data can be found in Hazlett, et al. (2017).

The primary wathershed moment of interest was the initiation of walking and its influence on the development of social behavior. The general hypothesis tested was that the timing to walking initiation would impact social behavior scores differently among EL-ASD infants from either EL-nonASD infants and LL-nonASD infants. Data from IBIS at the 6-, 12-, and 24-month visits were used to test this hypothesis. Sex and ASD likelihood group were required for a participant to be included in the analyses. The Vineland Adaptive Behavior Scale (VABS) was used to determine social behavior scores, and the standard socialization score was used in the analyses (Sparrow & Cicchetti, 1989). Fractional anisotropy (FA) values of fiber tracts from the MRI scans were used to measure development of occipital tracts. Specifically, the left and right inferior fronto-occipital fasciculus (IFOF) and left and right inferior longitudinal fasciculus (ILF) were examined. Initiation of walking was determined from the revised Autism Diagnostic Interview (ADI-R) age of first walking question, which asked parents to recall the age at which their infant first walked (in months).

The MRI scan data were during natural sleep using identical 3-T Siemens TIM Trio scanners equipped with 12-channel head coils at the four clinical sites. The diffusion tensor imaging sequence was acquired with a field of view of 190 mm (6 and 12 months) or 209 mm (24 months). Scan times ran 5 to 6 minutes. A total of 75–81 transversal slices with a slice thickness of 2 mm were collected using a 2×2×2-mm^3 voxel resolution, a repetition time of 12,800–13,300 ms, an echo time of 102 ms, and 25 gradient directions. Data from diffusion-weighted imaging were screened by using DTIprep software (Liu, et al., 2010). Label maps for the fiber tracts were generated in 3D Slicer (www.slicer.org; Kikinis, Pieper, & Vosburgh, 2014). FA values (range from 0, isotropic diffusion in fluid, to 1, strong directional diffusitivity) were generated

for each fiber tract. See Wolff, et al. (2012) and Girault, et al. (2022) for more information about the IBIS MRI and FA data collection processes.

Methods

The goal of these analyses was to determine whether the initiation of walking in an infant had cascading effects on social behavior and whether that downstream effect differentiated between ASD likelihood groups. As these ASD likelihood groups likely have heterogeneous profiles (Prince & Fidler, 2021; Bradshaw, Schwichtenberg, & Iverson, 2022), the hypothesis tests motivated the use of a JLCMM with a HARE determined hazards model. The VABS standard socialization score (VABS SS) was used to represent a measure of social behavior. Change in the VABS SS was estimated using a linear mixed effects model. The time-to-event was estimated using the ADI-R age of first walking question, where age is measured in months. The question was asked during the 24-month visit as part of the behavioral assessment. The time-to-event model was estimated using HARE, where potential covariates of interest included sex, ASD likelihood group, an indicator of whether the child had been crawling by 6-months of age, and the left IFOF FA value.

The VABS SS covers play and leisure time, interpersonal relationships, and coping skills of the child, and the domain has been demonstrated to have various positive qualities such as validity, reliability, and concordance with other socialization measures (Farmer, et al., 2020). It has an average score of 100 with a standard deviation of 15 points. Lower scores indicate a more problematic behavior, so a higher score indicates better socialization skills.

Time-to-initiation of walking was defined as the time in months reported by parents that the child first began to walk for the ADI-R walking initiation question (Lord, Rutter, & Le Couteur, 1994). Because of the IBIS sampling structure, the ADI-R is not given equally to the three ASD likelihood groups with notably fewer LL-nonASD infants receiving this question. As such, this item was censored for a number of LL-nonASD participants. However, the degree of walking was established by a VABS motor domain question asking the degree to which a child walks (“sometimes” or “usually”) at the 12-month and 24-month visits. Therefore, walking

status for LL-nonASD infants at specific visits. Figure 4.4.1 illustrates Kaplan-Meier curves demonstrating the survival probability of not walking for these three groups. Because of the measurement non-invariance of this walking measure, LL-nonASD kids have a larger survival probability than either EL groups. This non-invariance posed an interesting justification for the use of a JLCMM. The effect of likelihood group could be estimated within latent classes in order to mitigate the measurement non-invariance posed by the time-to-event outcome. Table 4.4.1 shows summary statistics for the independent and dependent variables used in the JLCMM broken down by ASD likelihood group.

The longitudinal model for this analysis was

$$y_i(t) = \beta_{0k} + b_{i0} + \beta_{1k}x_{i1} + \beta_{2k}x_{i2} + \beta_{3k}x_{i3} + (\beta_4 + b_{i1})x_{i4} + \beta_5 x_{i5} + \epsilon_i(t),$$

where $y_i(t)$ is the VABS SS for participant i at visit t, $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \beta_{2k}, \beta_{3k})$ is the vector of class-specific fixed-effects estimates for latent class k, $\boldsymbol{\beta} = (\beta_4, \beta_5)$ is the vector of marginal fixed-effects estimates, $\mathbf{b} = (b_{i0}, b_{i1})$ is a vector of random-effects, and $\epsilon_i(t)$ is the random error for participant i at visit t. The independent variables x_{i1}, x_{i2}, and x_{i3} were estimated as class-specific effects. Variable x_{i1} was the ASD likelihood group with three categories (EL-ASD, EL-Neg, LL-Neg) where EL-ASD infants were the referent group. Variable x_{i2} was a sex variable with female being the referent group. Variable x_{i3} was a binary indicator of whether the infant was crawling by 6-months of age with 'No' as the referent group. The variables x_{i4} and x_{i5} were estimated marginally rather than by each class k. Variable x_{i4} was the visit in months, and x_{i5} was was the FA value for the left IFOF. Left IFOF was selected to represent the FA tractography of the occipital lobe since the four tracts were highly collinear ($\rho_{min} > 0.75$). The random-effect b_{i0} was a random intercept for each participant i, and b_{i1} was a random slope for visit. The survival model was estimated as

$$\lambda_i(t) = \lambda_{0k}(t) \exp\{\theta_{1k}x_{i1} + \theta_{2k}x_{i2} + \theta_{3k}x_{i3} + \theta_4 x_{i5}\}.$$

The independent variables in the survival model are identical to the ones in the survival model, with the exception of no visit variable x_{i4} in the survival model. The vector $\boldsymbol{\theta} = (\theta_{1k}, \theta_{2k}, \theta_{3k}$ is a vector of class-specific estimates in the survival model. θ_4 is a marginal effect in the survival model. The survival model determined by HARE included knots at 6-months and 8-months-post first visit of (12-to-14-months of age) in addition to these independent variables as influential in the survival model. HARE was performed using the `hare()` function in the **polspline** package in R v4.1.0 (Kooperberg, 2020). Latent class membership was modeled as

$$P(c_i = k|x_{i1}) = \frac{\exp\left(\xi_{0k} + \xi_{1k}x_{i1}\right)}{\sum_{g=1}^{K}\exp\left(\xi_{0g} + \xi_{1g}x_{i1}\right)}$$

where ξ_{0k} is an intercept term and ξ_{1k} is the coefficient for ASD likelihood group.

Models with 1 - 4 latent classes were tested. The JLCMMs were estimated using the `Jointlcmm()` function from the **lcmm** package in R v4.1.0 (Proust-Lima, et al., 2017). All models used the `hazard = ''Specific''` option and `hazardnodes = c(6,8)` option to correspond to the knots in time. Parameters were estimated using the `gridsearch()` function with the one-class solution as the initialization model, 30 repetitions used for each iteration, and a maximum of 15 iterations specified.

Results

Figure 4.4.1 illustrates that the censoring imposed measurement noninvariance in estimation of the initiation to walking outcome. It is unlikely that most LL-nonASD infants had not initiated walking by the 8-month time point, so this measurement error justifies looking within class-specific estimates of ASD likelihood group to see whether any likelihood group differences exist within a latent class. While this comparison will also likely be biased, it will give a

better indication about the differences in hazard among latent class members who match a heterogeneous profile of motor and social development.

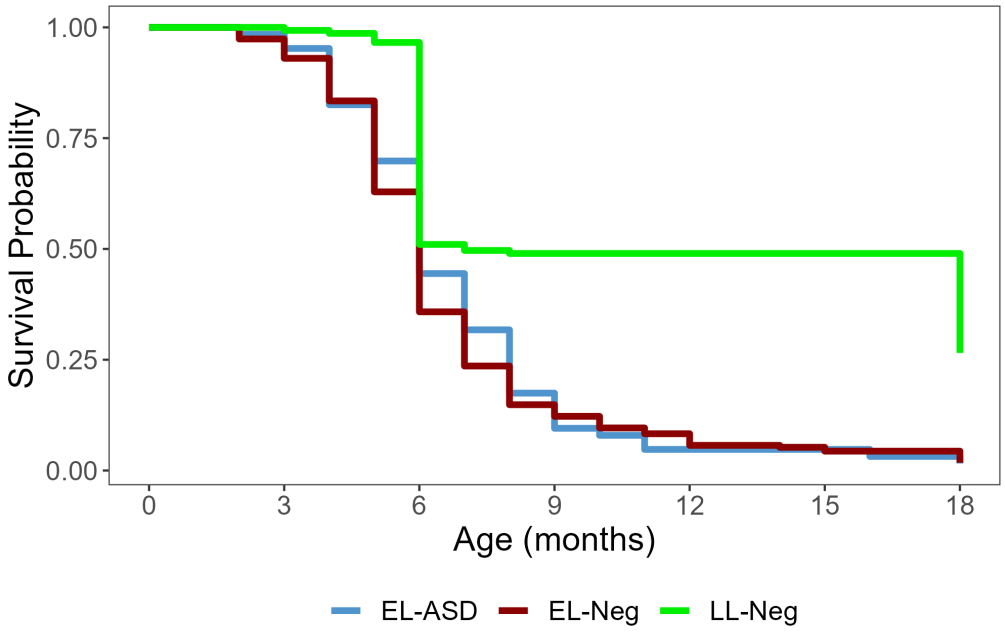

Figure 4.4.1 Kaplan-Meier Curves of Estimated Survival Probability of Not Walking for the Three ASD Likelihood Groups

Table 4.4.1 shows the characteristics of the sample divided into ASD likelihood group. The EL-ASD group is predominantly male, has a lower VABS SS score, and a smaller left IFOF at 6-months (compared to the LL-nonASD group). EL-nonASD infants have similar demographics to the EL-ASD infants, but their socialization scores are normal on average and are more evenly divided between boys and girls. The LL-nonASD group has the most developmentally advanced characteristics with larger left IFOF values, earlier walking initiation ages (despite the measurement error issues), and higher VABS socialization scores. This group also has a larger ratio of boys to girls, but not as dramatic a difference as the EL-ASD group.

Table 4.4.2 provides fit statistics and distributions of members for the one-to-four latent class JLCMMs. Both the two- and three-latent class solutions had reasonable fit with a BIC difference of around 6. While an argument could be made for selecting the model with two latent classes (Bauer, 2022), the model with three latent classes met the decision criterion

Table 4.4.1 Characteristics of the Sample by ASD Likelihood Group

Characteristic		EL-ASD (n = 63)	EL-Neg (n = 229)	LL-Neg (n = 147)
Sex				
	Male	53	125	87
	Female	10	104	60
Crawling				
	Yes	7	30	30
	No	56	199	117
Left IFOF (6 months)				
	Mean	0.369	0.369	0.373
	SD	0.024	0.024	0.021
	Range	0.325 - 0.424	0.308 - 0.425	0.320 - 0.455
Walking Initiation Age				
	Mean	12.5	12.2	11.6
	SD	2.28	2.25	1.59
	Range	8 - 22	8 - 21	9 - 14
VABS SS (6 months)				
	Mean	96.8	99.8	102
	SD	12.1	10.5	10.5
	Range	62 - 123	70 - 132	76 - 132

and was selected as the optimal model. The four class solution would not converge onto a maximum likelihood with four separate classes. Tables 4.4.3 and 4.4.4 provide details about the discriminatory power of the posterior probabilities used in the latent class assignment. The mean posterior probabilities for assignment into each latent class was over 85% with LC 2 having a mean posterior probability of assignment close to 100%. LC 3 had a small mean posterior probability of assignment to LC 1 (0.118), and LC 1 had an even smaller mean posterior probability of assignment to LC 3 (0.073). The proportion of posterior probability assignments ≥ 0.9 were all above 60% for each latent class with LC 2 having 100% of its probabilities over 0.9. These results indicated a very discriminatory assignment of the latent classes with each group being distinct from the other. Figure 4.4.2 illustrates the predicted VABS SS score for each of the three latent classes overlayed with individual trajectories. The consistency of the VABS SS over time can be seen, and two groups from these trajectories are apparent. One group of participants score around average throughout the study, and the other group score below average. This below-average group comprises mostly LC 1 and LC 3 with a few members from LC 2. Among the whole sample, mean VABS SS for LC 2 is the highest followed by LC 1 and LC 3, respectively. Figure 4.4.3 illustrates the hazard function of walking initiation over age in months for these three latent classes. All three latent classes have similar hazards for the first five months until LC 3 begins to increase its hazard of walking initiation. Around 10-months of age LC 1 has an exponential hazard increase in the initiation of walking. LC 2 has a small (but non-zero) hazard from 5-months of age and onward. Figure 4.4.4 illustrates the survival probability of not walking over age in months for the three latent classes. These curves parallel the results from the hazard function. LC 3 has the smallest curve of the three latent classes with its members initiating walking earliest and nearly all members having initiated this watershed moment by 10-months of age. LC 1 has a similar, but later trajectory compared to LC 3 where most members initiate walking by 12-months of age. LC 2 has the largest survival

curve indicating non-walking of members well beyond the ages of those from LC 1 or LC 3 (likely due to censoring).

Table 4.4.2 Joint Latent Class Mixed Model Summaries of Cascading Effects Models

K	LL	npm	BIC	% Class 1	% Class 2	% Class 3	% Class 4
1	-4055.1	22	8241.7	100%			
2	-3961.5	28	8162.2	60.4%	39.6%		
3	-3994.5	40	8156.5	74.2%	16.7%	9.1%	
4	-3979.2	52	8269.4	10.9%	34.6%	0%	54.5%

Table 4.4.3 Mean of Posterior Probabilities for the Three Latent Class Model

	$P(c_i = 1)$	$P(c_i = 2)$	$P(c_i = 3)$
1 (n = 294)	0.927	0.000	0.073
2 (n = 66)	0.000	0.999	0.000
3 (n = 36)	0.118	0.000	0.882

Table 4.4.4 Proportion of Posterior Probabilities above Various Thresholds

	k = 1 (n = 294)	k = 2 (n = 66)	k = 3 (n = 36)
$P(c_i = k) \geq 0.7$	96.3%	100%	86.1%
$P(c_i = k) \geq 0.8$	87.1%	100%	72.2%
$P(c_i = k) \geq 0.9$	77.2%	100%	61.1%

Characteristics of the three latent classes are given in Table 4.4.5. There did not appear to be any latent class that favored one sex over the other, perhaps excepting LC 1 having 1.6 times as many boys as girls, whereas for LC 2 and LC 3 these ratios are 1.2 and 1.4, respectively. LC 3 favored assignment from those who were not crawling by 6-months of age, which was also the group with the lowest mean VABS SS trajectories. LC 2 was primarily composed of

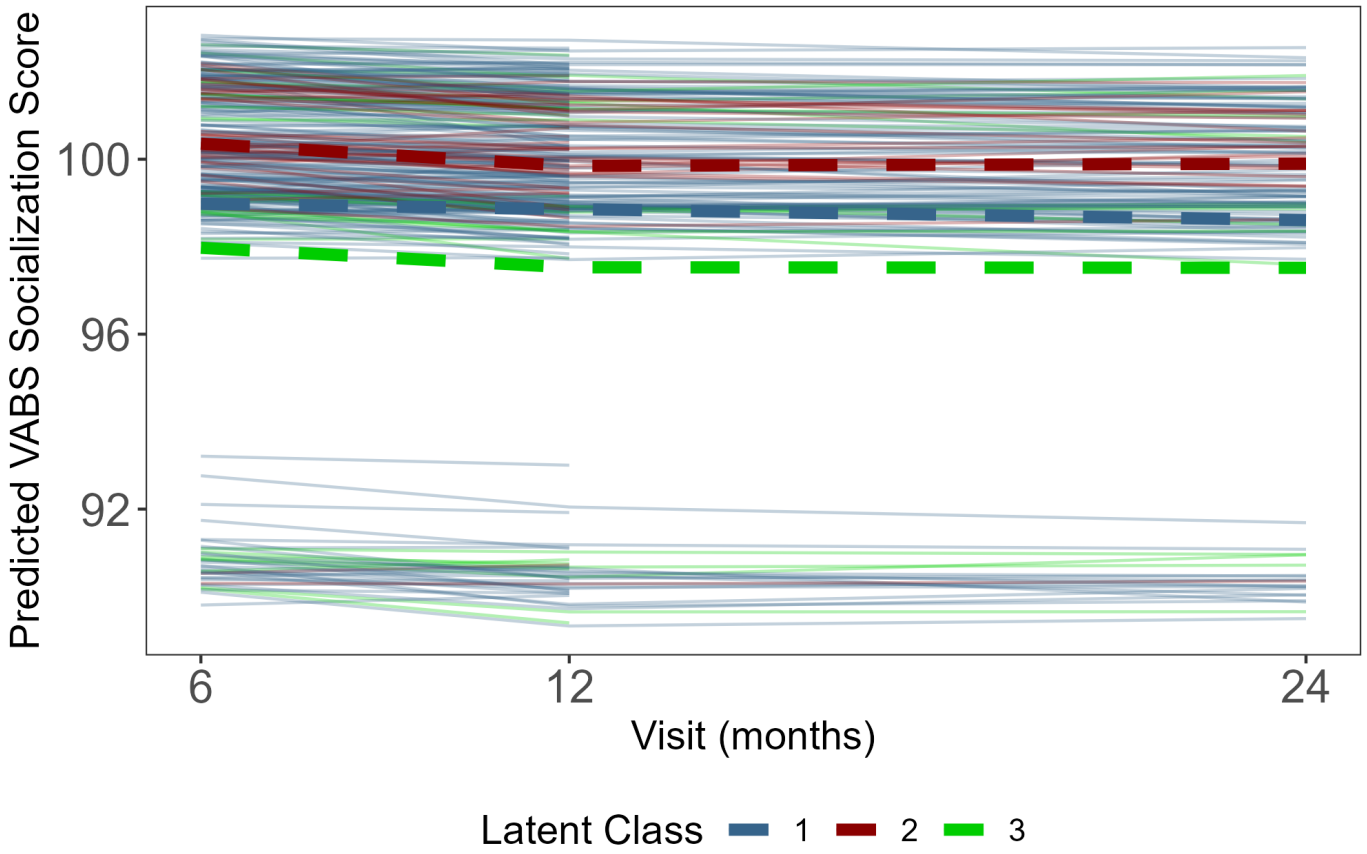

Figure 4.4.2 JLCMM Model-Predicted VABS Socialization Score

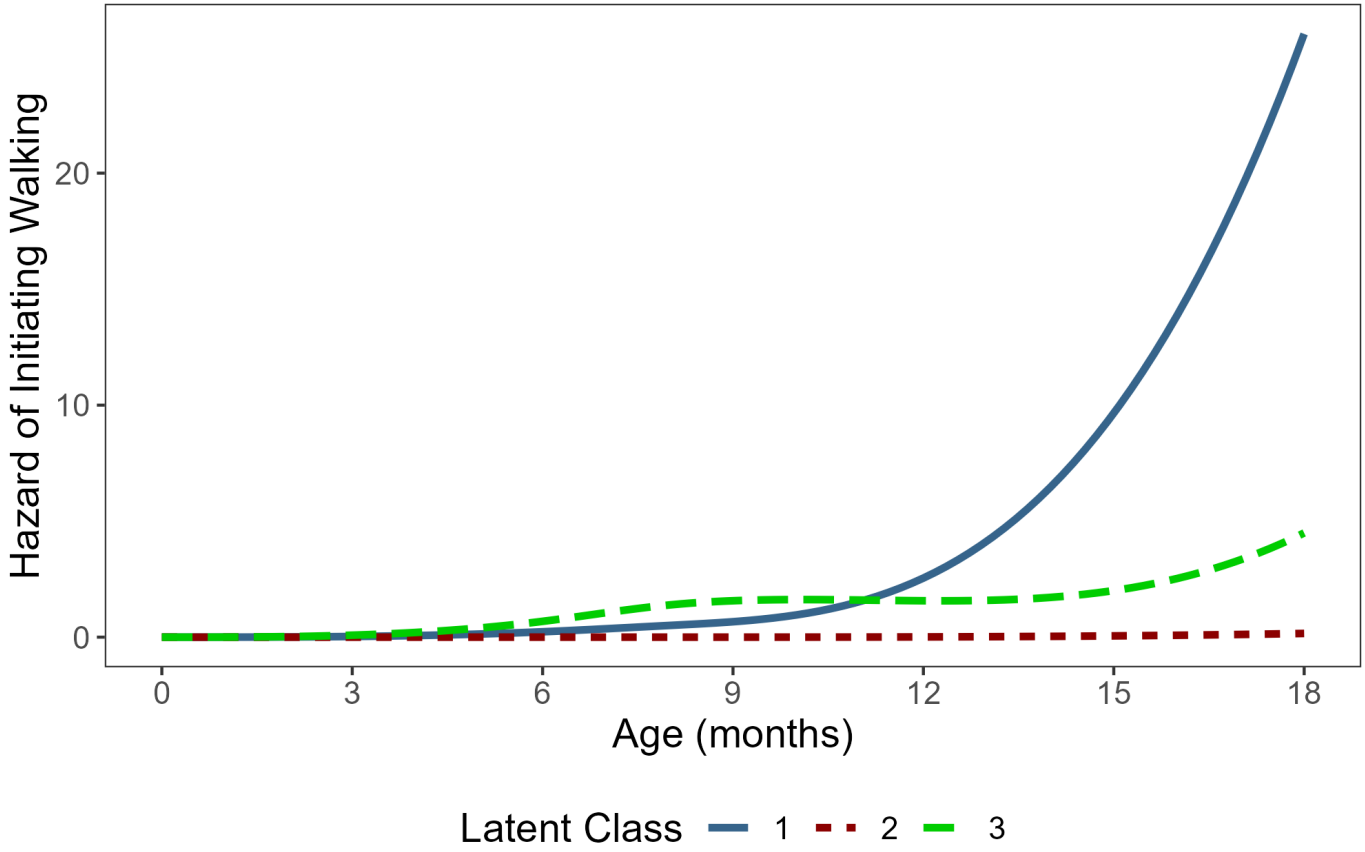

Figure 4.4.3 Hazard Function of Walking Initiation in months from Three-Class JLCMM

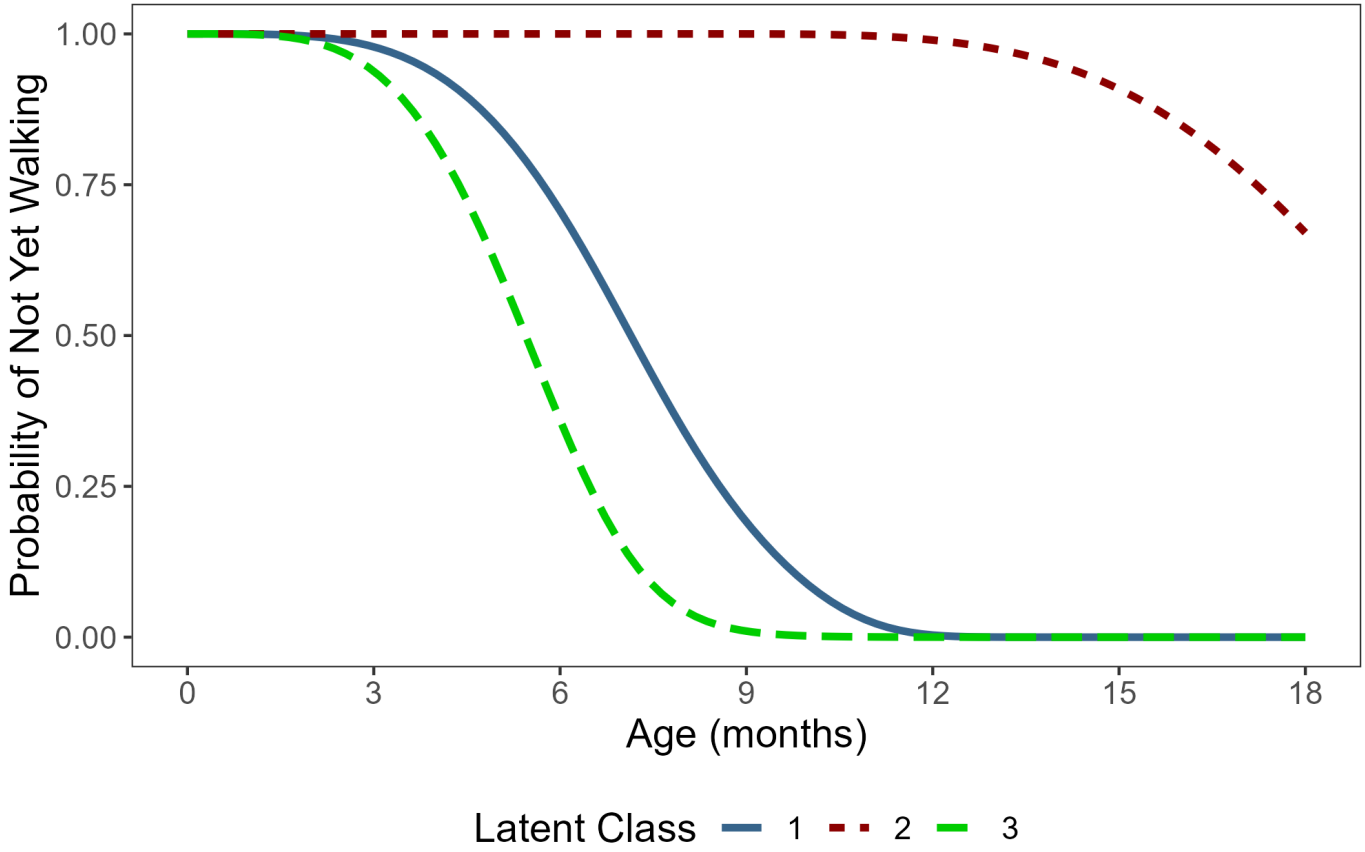

Figure 4.4.4 Probability of Not Walking in months from Three-Class JLCMM

LL-negative kids, and in particular had a large proportion of those kids with censored walking initiation data as reflected from Figure 4.4.1.

Table 4.4.5 Characteristics of the Latent Classes by ASD Likelihood Group

Characteristic		LC 1 (n = 294)	LC 2 (n = 66)	LC 3 (n = 36)
Sex				
	Male	180	36	21
	Female	114	30	15
Crawling				
	Yes	45	13	3
	No	249	53	33
ASD Likelihood Group				
	EL-ASD	44	3	9
	EL-Neg	180	0	27
	LL-Neg	70	63	0
Censored at last visit?				
Yes				
	EL-ASD	0	1	0
	EL-Neg	0	0	3
	LL-Neg	0	33	0
No				
	EL-ASD	44	2	9
	EL-Neg	180	0	24
	LL-Neg	70	30	0

Results from the three-class JLCMM model are presented in Table 4.4.6. Class-specific estimates of the baseline hazard function were excluded from this table to focus on hypothesis testing. Recall that the largest latent class (k = 3) is the referent group for the latent class membership model. Within LC 1, the average performing latent class, EL-nonASD participants had a higher hazard of initiating walking compared to their EL-ASD counterparts (HR = $e^{0.806}$ = 2.24, p-value = 0.0001) as did LL-nonASD participants (HR = $e^{0.747}$ = 2.11, p-value = 0.0005). However, among LC 1–the early-watershed/low socialization group–EL-nonASD participants had a lower hazard compared to their EL-ASD counterparts of initiating walking ($e^{-2.461}$ = 0.085, p-value < 0.0001). LL-nonASD participants did not have any statistically significant differences from EL-ASD participants in any LC other than LC 1. Note that the standard errors for EL-nonASD in LC 2 and LL-nonASD participants in LC 3 are exaggeratedly high in the

survival model. This problem occurs due to the absence of these members within the respective latent classes. Crawling at 6-months indicated a higher hazard of initiating walking in LC 1 ($e^{0.818}$ = 2.27, p-value < 0.0001). No other survival model results were statistically significant. Among the results for the linear mixed-effects model, there were statistically significantly higher VABS socialization scores for EL-nonASD participants ($\beta = 10.23$, p-value < 0.0001) and LL-nonASD participants ($\beta = 11.79$, p-value < 0.0001) compared to EL-ASD participants in LC 1. Positive differences occurred in LC 2 and LC 3 among these comparisons, but none were statistically significantly different. These differences indicate that EL-nonASD and LL-nonASD participants had better social behavior compared to EL-ASD participants in the average performing latent class. The only other result of note was that boys within LC 1 had lower VABS socialization scores compared to girls ($\beta = -2.516$, p-value < 0.0001).

The three-class JLCMM results indicated that social behavior and initiation to walking were inversely related: earlier initiation of walking, as measured by the month at which children began walking, was associated with worse social behavior outcomes. However, EL-nonASD and LL-nonASD participants had better social behavior outcomes and earlier walking initiations compared to their EL-ASD counterparts. Therefore, these results have one minor problem: they make absolutely no sense. From a substantive standpoint, evidence already exists in the literature that indicate earlier initiation of movement and walking lead to higher levels of social behavior (Smith, et al., 2018). From a statistical standpoint, a situation where the association within a cluster behaves antithetically to the overall trend reeks of Simpson's paradox (Simpson, 1951). While the measurement noninvariance of walking initiation may have been the sole issue, model misspecification could have exacerbated the issue. One natural modification to the model was to estimate class-specific estimates for the left IFOF value in the survival and longitudinal models (i.e., $\beta_5 \rightarrow \beta_{5k}$ and $\theta_4 \rightarrow \theta_{4k}$). Latent class models with 1 - 4 latent classes were rerun. In order to reach convergence, the maximum number of iterations had to be raised to `maxiter = 50`. Tables 4.4.7 - 4.4.11 provide identical results as 4.4.2 - 4.4.6 using the modified model. Figures 4.4.5 - 4.4.7 provide identical results as 4.4.2 - 4.4.4 using the modified model.

Table 4.4.6 Estimates of Coefficients from the JLCMM

Class Membership Model				
Parameter	ξ	SE	t-value	p-value
Intercept LC 1	1.133	0.875	1.294	0.196
Intercept LC 2	-1.444	0.889	-1.623	0.105
EL-Negative LC 1	0.289	0.927	0.311	0.755
EL-Negative LC 2	-13.702	594.7	-0.023	0.982
LL-Negative LC 1	9.109	48.27	0.189	0.850
LL-Negative LC 2	11.581	48.26	0.240	0.810
Survival Model				
Parameter	θ	SE	t-value	p-value
EL-Negative LC 1	0.806	0.208	3.866	0.0001
EL-Negative LC 2	1.550	274.4	0.006	0.995
EL-Negative LC 3	-2.461	0.313	-7.867	< 0.0001
LL-Negative LC 1	0.747	0.216	3.460	0.0005
LL-Negative LC 2	-0.358	0.377	-0.950	0.342
LL-Negative LC 3	-7.267	460.9	-0.016	0.987
Male LC 1	-0.016	0.139	-0.120	0.905
Male LC 2	0.652	0.356	1.829	0.067
Male LC 3	0.275	0.314	0.876	0.381
Crawling at 6-months LC 1	0.818	0.181	4.526	< 0.0001
Crawling at 6-months LC 2	0.207	0.476	0.435	0.664
Crawling at 6-months LC 3	-1.152	0.639	-1.803	0.0714
Left IFOF	0.263	0.551	0.478	0.633
Mixed-Effects Model				
Parameter	β	SE	t-value	p-value
Intercept LC 1	97.23	5.63	17.29	< 0.0001
Intercept LC 2	99.08	7.48	13.25	< 0.0001
Intercept LC 3	102.76	7.13	14.42	< 0.0001
Visit	0.041	0.077	0.535	0.592
EL-Negative LC 1	10.23	1.824	5.610	< 0.0001
EL-Negative LC 2	5.99	511.50	0.012	0.991
EL-Negative LC 3	0.219	4.531	0.048	0.961
LL-Negative LC 1	11.79	1.948	6.050	< 0.0001
LL-Negative LC 2	7.520	4.818	0.020	0.984
LL-Negative LC 3	0.034	1.711	0.020	0.119
Male LC 1	-2.516	1.063	-2.367	0.018
Male LC 2	-1.467	2.178	-0.673	0.501
Male LC 3	0.929	2.952	0.315	0.753

Parameter (cont'd)	θ	SE	t-value	p-value
Crawling at 6-months LC 1	-0.696	1.324	-0.525	0.599
Crawling at 6-months LC 2	1.941	2.756	0.704	0.481
Crawling at 6-months LC 3	0.666	4.379	0.152	0.879
Left IFOF	-16.08	15.05	-1.069	0.285
Σ	Intercept	Visit		
Intercept	45.06			
Visit	-1.593	0.093		

Fit statistics for the modified model resulted in the selection of a three-class model (BIC = 8120.4). The BIC for this model was better than for the initial three-class model. This class consolidates more individuals into one latent class (LC 2) than assigned in the original model. Discrimination of these latent classes were also fairly high with large assignment posterior probabilities for all three latent classes (Tables 4.4.8 and 4.4.9).

Table 4.4.7 Joint Latent Class Mixed Model Summaries of Modified Cascading Effects Models

K	LL	npm	BIC	% Class 1	% Class 2	% Class 3	% Class 4
1	-4055.1	22	8241.7	100%			
2	-3948.1	30	8159.4	79.8%	20.2%		
3	-3970.3	44	8120.4	19.7%	75.5%	4.8%	
4	-3953.4	58	8253.6	10.9%	34.6%	0%	54.5%

Table 4.4.8 Mean of Posterior Probabilities for the Modified Three Latent Class Model

	$P(c_i = 1)$	$P(c_i = 2)$	$P(c_i = 3)$
1 (n = 78)	0.998	0.002	0.000
2 (n = 299)	0.008	0.911	0.082
3 (n = 19)	0.000	0.155	0.845

Table 4.4.9 Proportion of Posterior Probabilities above Various Thresholds for Modified JLCMM

	k = 1 (n = 78)	k = 2 (n = 299)	k = 3 (n = 19)
$P(c_i = k) \geq 0.7$	100%	93.0%	84.2%
$P(c_i = k) \geq 0.8$	100%	80.6%	68.4%
$P(c_i = k) \geq 0.9$	100%	65.2%	47.4%

Table 4.4.10 Characteristics of the Modified Latent Classes by ASD Likelihood Group

Characteristic		LC 1 (n = 78)	LC 2 (n = 299)	LC 3 (n = 19)
Sex				
	Male	42	180	15
	Female	36	119	4
Crawling				
	Yes	16	42	3
	No	62	257	16
ASD Likelihood Group				
	EL-ASD	3	46	7
	EL-Neg	12	184	11
	LL-Neg	63	69	1
Censored at last visit?				
Yes				
	EL-ASD	1	0	0
	EL-Neg	3	0	0
	LL-Neg	33	0	0
No				
	EL-ASD	2	46	7
	EL-Neg	9	184	11
	LL-Neg	30	69	1

The predicted trajectories, hazard functions, and survival plots are illustrated in Figures 4.4.5 - 4.4.7. The patterns of the joint outcomes indicate three groups: LC 1 has an above-average socialization score over time with a later initiation to walking, LC 2 has an average socialization score over time with an initiation to walking that occurs primarily between 6-to-12-months of age, and LC 3 has a below average socialization score that improves over time with a later initiation to walking that eventually catches up to LC 2. LC 1 still has an issue of inversely related outcomes, likely related to the proportion of LL-nonASD members in this class, but the LC behaves more realistically than LC 2 in the first model. Further, there is better mixing between latent classes and ASD likelihood groups with the modified model. LC 1, the high socialization and slow initiation class, did have a preponderance of LL-nonASD members who

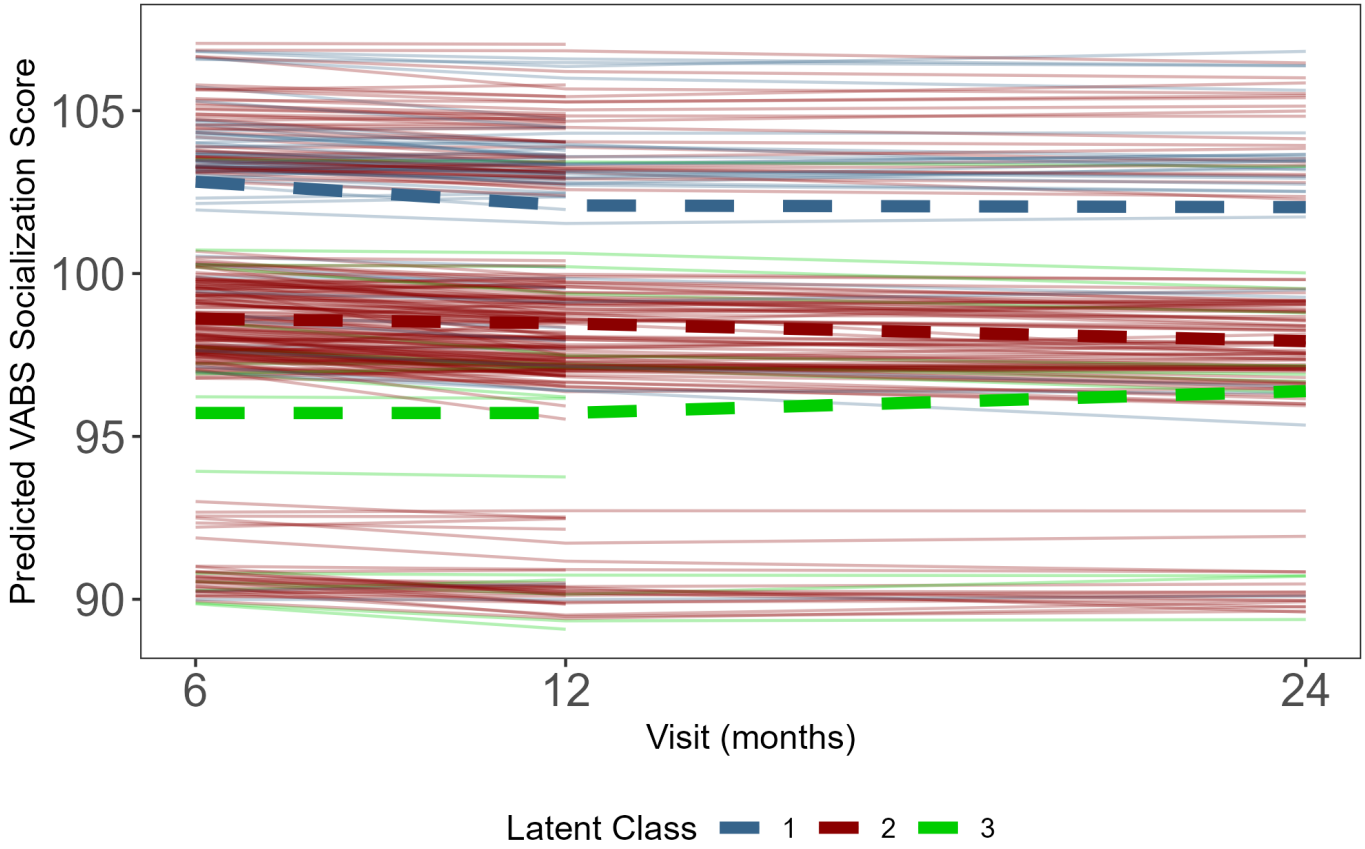

Figure 4.4.5 Modified JLCMM Model-Predicted VABS Socialization Score over Three Visits

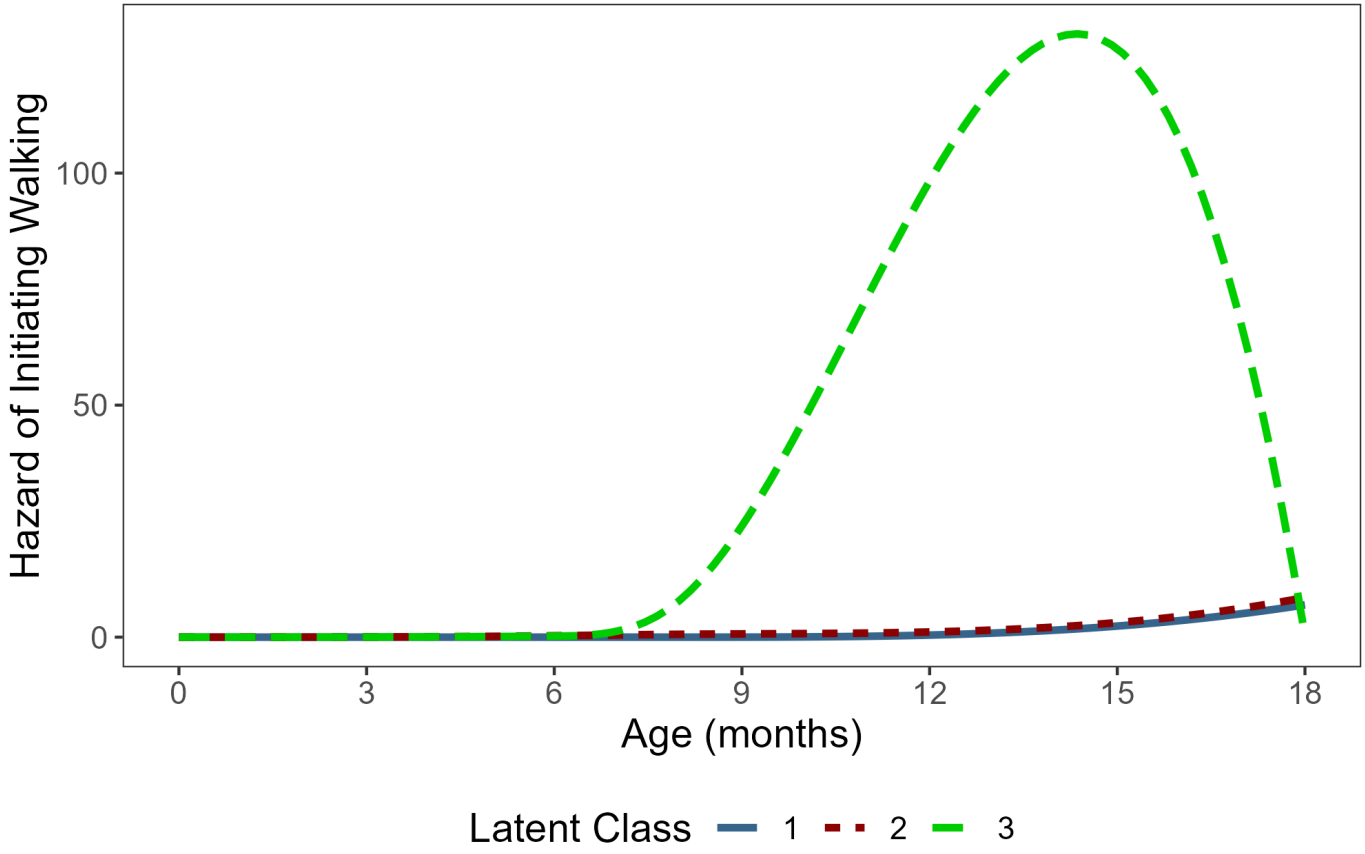

Figure 4.4.6 Hazard Function of Walking Initiation in months from Modified Three-Class JLCMM

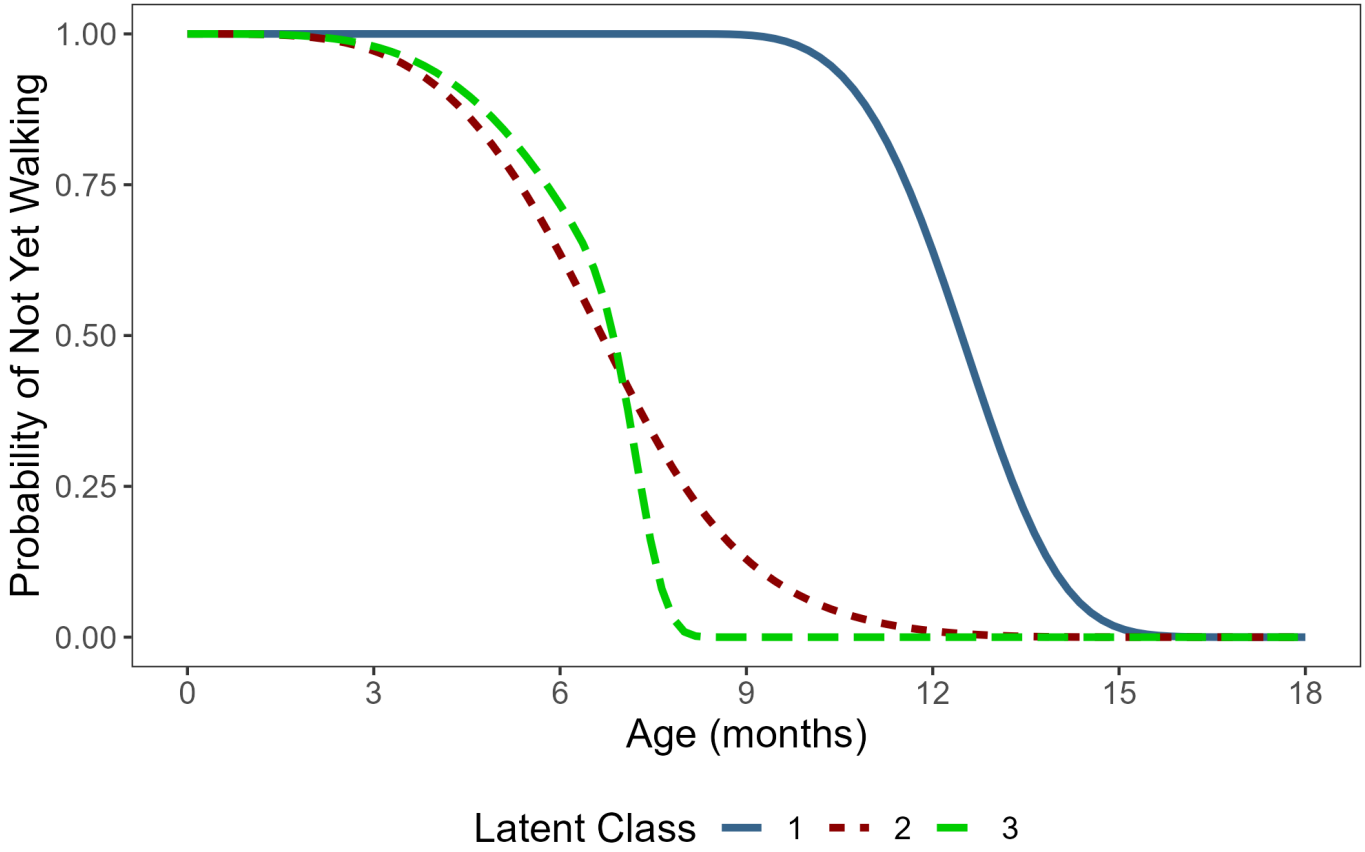

Figure 4.4.7 Probability of Not Walking in months from Modified Three-Class JLCMM

were censored at the last visit in the class. This likely affected the measurement of walking initiation even in the modified model.

Results from the modified three-class JLCMM model are presented in Table 4.4.11. Class-specific estimates of the baseline hazard function were excluded from this table to focus on hypothesis testing. Recall that the largest latent class (k = 3) is the referent group for the latent class membership model. Within LC 2, the average performing latent class, though not statistically significant, EL-nonASD participants had a higher hazard of initiating walking compared to their EL-ASD counterparts (HR = $e^{0.361}$ = 1.43, p-value = 0.068) as did LL-nonASD participants (HR = $e^{0.679}$ = 1.97, p-value = 0.001). No other statistically significant differences of hazard between ASD likelihood groups were found among the classes. Crawling at 6-months indicated a higher hazard of initiating walking in LC 2 ($e^{0.881}$ = 2.41, p-value < 0.0001) and LC 3 ($e^{2.01}$ = 7.46, p-value= 0.006). Further, the left IFOF values were negatively associated with initiation to walking in LC 1, but positively associated in LC 3. This conflict likely occurs because of the larger average left IFOF value among LL-nonASD participants (Table 4.4.1) coupled with the fact that LC 1 comprised mostly LL-nonASD kids, half of whom were censored from walking initiation. No other survival model results were statistically significant. Among the results for the linear mixed-effects model, there were statistically significantly higher VABS socialization scores for EL-nonASD participants and LL-nonASD participants compared to EL-ASD participants in all three latent classes, excepting EL-nonASD in LC 1. Boys had worse socialization outcomes than girls in LC 2, but not in LC 3. However, only 4 girls were assigned to LC 3. Crawling at 6-months was positively associated with VABS SS ($\beta = 22.05$, p-value = 0.0001). This result, while statistically significant, only affects three participants of the 19 in LC 3. No other longitudinal model results were statistically significant.

Table 4.4.11 Estimates of Coefficients from the Modified JLCMM

Class Membership Model				
Parameter	ξ	SE	t-value	p-value
Intercept LC 1	-1.186	0.762	-1.557	0.120
Intercept LC 2	1.460	0.590	2.474	0.013
EL-Negative LC 1	0.477	0.907	0.526	0.599
EL-Negative LC 2	0.279	0.732	0.381	0.703
LL-Negative LC 1	4.825	1.311	3.680	0.0002
LL-Negative LC 2	2.260	1.228	1.840	0.066
Survival Model				
Parameter	θ	SE	t-value	p-value
EL-Negative LC 1	0.461	0.751	0.615	0.539
EL-Negative LC 2	0.361	0.198	1.823	0.068
EL-Negative LC 3	-0.283	0.488	-0.579	0.562
LL-Negative LC 1	-0.347	0.709	-0.490	0.624
LL-Negative LC 2	0.679	0.206	3.288	0.001
LL-Negative LC 3	-0.342	0.872	-0.392	0.695
Male LC 1	0.361	0.332	1.087	0.277
Male LC 2	0.098	0.139	0.700	0.483
Male LC 3	0.197	0.487	0.404	0.686
Crawling at 6-months LC 1	-0.156	0.413	-0.378	0.705
Crawling at 6-months LC 2	0.881	0.183	4.815	< 0.0001
Crawling at 6-months LC 3	2.01	0.727	2.765	0.006
Left IFOF LC 1	-8.913	1.948	-4.576	< 0.0001
Left IFOF LC 2	-0.533	0.544	-0.979	0.327
Left IFOF LC 3	3.561	1.496	2.381	0.017
Mixed-Effects Model				
Parameter	β	SE	t-value	p-value
Intercept LC 1	97.39	9.21	10.57	< 0.0001
Intercept LC 2	102.66	6.57	15.62	< 0.0001
Intercept LC 3	68.52	16.15	4.24	< 0.0001
Visit	0.045	0.074	0.611	0.541
EL-Negative LC 1	5.74	4.46	1.288	0.541
EL-Negative LC 2	8.79	1.43	6.151	< 0.0001
EL-Negative LC 3	9.26	3.85	2.404	0.016
LL-Negative LC 1	8.10	4.13	1.962	0.050
LL-Negative LC 2	11.33	1.57	7.209	< 0.0001
LL-Negative LC 3	24.21	7.28	3.327	0.0009
Male LC 1	-0.774	1.668	-0.464	0.643
Male LC 2	-3.216	1.022	-3.147	0.002

Parameter (cont'd)	θ	SE	t-value	p-value
Male LC 3	9.589	4.210	2.278	0.023
Crawling at 6-months LC 1	1.964	2.079	0.945	0.345
Crawling at 6-months LC 2	-1.174	1.204	-0.975	0.329
Crawling at 6-months LC 3	22.05	5.678	3.885	0.0001
Left IFOF LC 1	-14.64	20.97	-0.698	0.485
Left IFOF LC 2	-28.72	17.04	-1.685	0.092
Left IFOF LC 3	52.17	39.53	1.320	0.187
Σ	Intercept	Visit		
Intercept	41.17			
Visit	-1.892	0.087		

Discussion

While some measurement noninvariance of the walking initiation variable still posed some identification issues, the modified JLCMM improved the results of the original model. Performing hypothesis tests between the ASD likelihood groups became feasible after the modification because the ASD likelihood groups were not strictly divided within different latent classes. However, it was still important to produce these latent classes in order to separate those who had noninvariant walking initiation ages from participants who had more reliable estimates. HARE determined two knots in time for the hazard model at 6-months and 8-months after study initiation, roughly equating to 12-months and 14-months of age. While a twelve month knot could likely have been determined not using a data-driven method, the 14-month determination would not have been immediately apparent either from using quantiles or cursory data summary methods. Using only 12-months of age/a 6-month time knot produced poorly fitting results (not shown). HARE was able to determine important parts of the sequence of walking initiation months to estimate the hazard function that helped in distinguishing these latent classes. Future analyses ought to be conducted, however, with more complete walking initiation data.

The model improved in both fit and substantive validation after left IFOF values were allowed to vary by class. The regression results for these values were not statistically significant, but it appeared as if this tract mediated or affected the relationship between the two outcomes in some way. Further analyses need to be completed in order to understand the appropriate method

of incorporating tractography data into these analyses. The analyses conducted illustrated an apparent issue with latent class identification, where slight model misspecification and measurement error posed results incongruent with known phenomena. These problems become exacerbated since the JLCMM can fail to converge using spline-based baseline hazard functions, with HARE being no exception. These issues highlight the need for external validation of latent classification, even outside of the joint modeling and HARE framework.

CHAPTER 5: CONCLUSIONS

The analyses indicated that hazard regression (HARE) provided a novel method of estimating class-specific hazard functions within a joint latent class mixed model (JLCMM) framework because of its ability to adapt to heterogeneous profiles of survival. The novelty transititons into utility when one considers that HARE does not have the proportionality assumption inherent to Cox proportional hazards models. While some proportionality between hazards may exist, such as demonstrated in the Paquid data analysis, a notable amount of social science data indicate non-proportional hazards of experiencing times-to-event whether through the inherent nature of the problem, through a confluence of related factors, or due to some measurement error. Looser assumptions that provide comparable or more realistic inferences become necessary as data stop conforming to probability theorems.

Results from the simulation study indicated that when the model is well-specified, such as being simulated from a parametric distribution, HARE provided comparable estimates to a survival model that assumed a mixture of these parametric distributions. Further, the estimated effects had favorable asymptotic properties for larger datasets. The ability for the JLCMM to converge on the true global maximum likelihood did suffer using HARE unless effort was made to use optimal starting values. However, this limitation exists within many JLCMMs (Proust-Lima, et al., 2017) and is not more prevalent with HARE than any other spline-based method of knot determination. A more pressing concern was the larger bias in the hazard model estimates than the longitudinal outcome estimates. While these larger biases did parallel the ones from the Weibull model, one would hope that a flexible hazard function estimation method would produce less bias at the expense of more error. More research needs to be done to ascertain the reason for this issue.

The analysis of the social fixation arousal experiment data revealed the justification of both JLCMM use and the HARE method of hazard estimation within it. As shown by the post-hoc Cox models, using joint outcomes to estimate latent classes among a heterogeneous sample provided better estimates of the association between the two vision tasks than using a single model with a time-varying covariate. Further, since the hazard functions were not proportional between the three latent classes, a data-driven method of estimating those hazards likely improved the process. The key limitation from these analyses was the convergence issue. Having nonproportional hazard functions implies that some knots either in time or a covariate will have little, if any, variation within a latent class. This issue led to the Levenberg-Marquardt algorithm to maximize the likelihood function along a boundary space. Fixing the estimates around these boundary cases did solve the convergence problem, but it could have also biased estimates in an unforseeable way. The latent classes have some external validity, but resolving this optimization method would be preferable for future research.

The identification issues of any latent class analysis, including with JLCMMs, became apparent in the cascading effects analyses. Any model that assumes an uknown mixture of random distributions will contend with the issue of identification (Masyn, 2013). This conundrum exemplified the justification of substantive and external validation of latent class assignment. Since the latent classes contradicted established science, it was likely that they were valid despite the high discrimination of the classes themselves. Further model building improved the classification of the model. Although the modified model still contended with measurement error, the within-class hypothesis tests converged with prior theory and research about the relationship between development watershed moments among those with neurodevelopmental disorders. The point of concern with using HARE in this scenario derives from the knot determination. These knots were determined at 6-months and 8-months after first visit onset (accordingly 12-months and 14-months of age). Both ages are reasonable knot locations, but the determination for these locations is unknown outside of understanding the pattern of recursive partitioning performed by HARE. HARE, just like many model-fitting methods, did not provide

substantive reasoning for these two knot locations. It was very likely that a knot at 12-months of age was determined because it was the earliest time determined for many infants with a low likelihood of developing Autism Spectrum Disorder (ASD). While no statistical method can provide such detail, this kind of adaptive method can easily lead to erroneous latent classification if the model is not examined with care.

While HARE provided novelty and utility to the estimation of the survival model within a JLCMM, improvements could be made to this methodology that would overcome some limitations. Primarily, knot selection needs to account for ample data support within each latent class. Currently, the method assumes suitable variance in the neighborhood of each knot equally among latent classes. However, this assumption is unrealistic, especially in cases where hazard functions are not proportional. A sequence of time in which hazard increases greatly for one latent class may have no change in the data for another. This discordance has led to certain knots within a specific class having no variance, which has led to the majority of convergence issues (from Hessian singularity) with the analyses in this manuscript. Various options currently exist that may be implemented to help with this task. K-nearest neighbor methods (Lloyd, 1982) have been used to locate dense areas of data. Stone (1977) used this method to produce weight functions for regression in a consistent, nonparametric fashion. This method or one similar could be used within HARE to determine class-specific weights for knot determination.

A secondary improvement would be to find an alternative method for estimating the maximum likelihood estimates of the parameters. The Levenberg-Marquardt method appears to have issues when working with splines (Proust-Lima, et al., 2017), and a nonparametric method of estimation could minimize the troubleshooting associated with this process. An adaptive method similar to multivariate adaptive regression splines (Friedman, 1991) could be implemented for the longitudinal outcome of JLCMM so that both outcomes are estimated adaptively and nonparametrically. One obstacle to this implementation would be standard error estimation that needs to account for the randomness of knot selection. As noted in Kooperberg, Stone, and Truong (1995a), the standard errors estimated by HARE are ‘merely suggestive’.

However, other nonparametric methods have been successfully developed to estimate a JLCMM (Wong, Zeng, & Lin, 2022), so this potential method may not be intractable.

REFERENCES

Aalen, O. (1976). Nonparametric inference in connection with multiple decrement models. *Scandinavian Journal of Statistics*, 15-27.

Aguinis, H., & Pierce, C. A. (1998). *Heterogeneity of error variance and the assessment of moderating effects of categorical variables: A conceptual review.* Sage Publications.

Ajdacic-Gross, V., Bechtiger, L., Rodgers, S., Müller, M., Kawohl, W., von Känel, R., Mutsch, M., Rössler, W., Seifritz, E., Castelao, E. and Strippoli, M.P.F., Vandeleur, C., Preisig, M., & Howell, P. (2018). Subtypes of stuttering determined by latent class analysis in two Swiss epidemiological surveys. *PloS one*, 13(8), e0198450.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723.

Alvares, D., & Rubio, F. J. (2021). A tractable bayesian joint model for longitudinal and survival data. *Statistics in Medicine*, 40(19), 4213-4229.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.)

Bai, D., Yip, B. H. K., Windham, G. C., Sourander, A., Francis, R., Yoffe, R., Glasson, E., Mahjani, B., Suominen, A., Leonard, H., Gissler, M., Buxbaum, J. D., Wong, K., Schendel, D., Kodesh, A., Breshnahan, M., Levine, S. Z., Parner, E. T., Hansen, S. N., . . ., & Sandin, S. (2019). Association of genetic and environmental factors with autism in a 5-country cohort. *JAMA Psychiatry (Chicago, Ill.)*, 76(10), 1035-1043.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173-1182.

Bauer, J. (2022). A primer to latent profile and latent class analysis. *Professional and Practice-Based Learning*, 33, 243-268.

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models: Generating Survival Times. *Statistics in Medicine*, 24(11), 1713-1723.

Betancur, C. (2011). Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain research*, 1380, 42-77.

Bollen, K. A. (1989). *Structural equations with latent variables* (1st ed.). Wiley.

Bottema-Beutel, K., Kapp, S. K., Lester, J. N., Sasson, N. J., & Hand, B. N. (2021). Avoiding ableist language: Suggestions for autism researchers. *Autism in Adulthood*, 3(1), 18-29.

Bradshaw, J., Schwichtenberg, A. J., & Iverson, J. M. (2022). Capturing the complexity of autism: Applying a developmental cascades framework. *Child Development Perspectives*, 16(1), 18-26.

Brady, M. J., Cella, D. F., Mo, F., Bonomi, A. E., Tulsky, D. S., Lloyd, S. R., Deasy, S., Cobleigh, M., & Shiomoto, G. (1997). Reliability and validity of the functional assessment of cancer therapy-breast quality-of-life instrument. *Journal of Clinical Oncology*, 15(3), 974-986.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*.

Brown, E. R., Ibrahim, J. G., & De Gruttola, V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61(1), 64-73.

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, 104(3), 396-404.

Catalona, W. J., Smith, D. S., & Ornstein, D. K. (1997). Prostate cancer detection in men with serum PSA concentrations of 2.6 to 4.0 ng/mL and benign prostate examination: Enhancement of specificity with free PSA measurements. *JAMA : The Journal of the American Medical Association*, 277(18), 1452-1455.

Chi, Y., & Ibrahim, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, 62(2), 432-445.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6, 5549-5632.

Commenges, D., Liquet, B., & Proust-Lima, C. (2012). Choice of prognostic estimators in joint models by estimating differences of expected conditional Kullback–Leibler risks. *Biometrics*, 68(2), 380-387.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187-202.

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

Crofts, A., Kelly, M. E., & Gibson, C. L. (2020). Imaging functional recovery following ischemic stroke: clinical and preclinical fMRI studies. *Journal of Neuroimaging*, 30(1), 5-14.

De Boor, C. (1986). *B(asic)-Spline Basics*. Mathematics Research Center, University of Wisconsin-Madison .

De Gruttola, V., Lange, N., & Dafni, U. (1991). Modeling the Progression of HIV Infection. *Journal of the American Statistical Association*, 86, 569-577.

De Gruttola, V., & Tu, X. M. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, 50(4), 1003-1014.

Des Jarlais, D. C., & Friedman, S. R. (1988). Needle sharing among IVDUs at risk for AIDS. *American Journal of Public Health (1971)*, 78(11), 1498-1499.

Diggle, P., Liang, K. Y., & Zeger, S. L. (1994). *Longitudinal data analysis*. New York: Oxford University Press, 5, 13.

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, 43(1), 49-93.

Dudley, W. N., Wickham, R., & Coombs, N. (2016). An introduction to survival statistics: Kaplan-Meier analysis. *Journal of the advanced practitioner in oncology*, 7(1), 91.

Dykens, E. M. (2003). Anxiety, fears, and phobias in persons with williams syndrome. *Developmental Neuropsychology*, 23(1-2), 291-316.

Eikeseth, S., Klintwall, L., Jahr, E., & Karlsson, P. (2012). Outcome for children with autism receiving early and intensive behavioral intervention in mainstream preschool and kindergarten settings. *Research in Autism Spectrum Disorders*, 6(2), 829-835.

Elashoff, R. M., Li, G., & Li, N. (2008). A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics*, 64(3), 762-771.

Eldevik, S., Hastings, R. P., Jahr, E., & Hughes, J. C. (2012). Outcomes of behavioral intervention for children with autism in mainstream pre-school settings. *Journal of Autism and Developmental Disorders*, 42(2), 210-220.

Farmer, C., Adedipe, D., Bal, V. H., Chlebowski, C., & Thurm, A. (2020). Concordance of the vineland adaptive behavior scales, second and third editions. *Journal of Intellectual Disability Research*, 64(1), 18-26.

Fahey, J. L., Taylor, J. M. G., Detels, R., Hoffman, B., Melmed, R., Nishanian, P., & Giorgi, J. (1990). The Prognostic Value of Cellular and Serologica Markers in Infection with HIV Type 1. *New England Journal of Medicine*, 322, 166-172.

Faucett, C. L., & Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statistics in Medicine*, 15(15), 1663-1685.

Fidler, D. J., Prince, M. A., Van Deusen, K., Esbensen, A. J., Thurman, A. J., Abbeduto, L., Patel, L., Mervis, C., Schworer, E.K., Lee, N.R., Edgin, J.O., Hepburn, S., Davis, S. & Daunhauer, L. A. (2022). Latent profiles of autism symptoms in children and adolescents with Down syndrome. *Journal of Intellectual Disability Research*, 66(3), 265-281.

Fischl, M. A., Parker, C. B., Pettinelli, C., Wulfsohn, M., Hirsch, M. S., Collier, A. C., Antoniskis, D., Ho, M., Richman, D. D., Fuchs, E., Merigan, T. C., Reichman, R. C., Gold, J., Steigbigel, N., Leoung, G. S., Rasheed, S., Tsiatis, A., & the AIDS Clinical Trials Group. (1990). A randomized controlled trial of a reduced daily dose of zidovudine in patients with the acquired immunodeficiency syndrome. *The New England Journal of Medicine*, 323(15), 1009-1014.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis*. John Wiley & Sons, Incorporated.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3), 189-198.

Friedman, J. H., & Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31(1), 3-21.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1), 1-67.

Garre, F. G., Zwinderman, A. H., Geskus, R. B., & Sijpkens, Y. W. J. (2008). Joint latent class changepoint model to improve the prediction of time to graft failure. *Journal of the Royal Statistical Society. Series A, Statistics in Society*, 171(1), 299-308.

Geier, D. A., & Geier, M. R. (2021). A longitudinal cohort study of precocious puberty and autism spectrum disorder. Hormone Research in Paediatrics, 94(5-6), 219-228.

Gelman, A. (2005). Analysis of variance—why it is more important than ever. *The annals of statistics*, 33(1), 1-53.

Girault, J. B., & Piven, J. (2020). The neurodevelopment of autism from infancy through toddlerhood. *Neuroimaging Clinics of North America*, 30(1), 97-114.

Girault, J. B., Donovan, K., Hawks, Z., Talovic, M., Forsen, E., Elison, J. T., Shen, M. D., Swanson, M. R., Wolff, J. J., Kim, S. H., Nishino, T., Davis, S., Snyder, A. Z., Botteron, K. N., Estes, A. M., Dager, S. R., Hazlett, H. C., Gerig, G., McKinstry, R., . . . for the IBIS Network. (2022). Infant visual brain development and inherited genetic liability in autism. *The American Journal of Psychiatry*, 179(8), 573-585.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, 68(1), 155–165.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, 75(5), 861-871.

Grzadzinski, R., Huerta, M., & Lord, C. (2013). DSM-5 and autism spectrum disorders (ASDs): An opportunity for identifying ASD subtypes. *Molecular Autism*, 4(1), 12-12.

Grzadzinski, R., Amso, D., Landa, R., Watson, L., Guralnick, M., Zwaigenbaum, L., Deák, G., Estes, A., Brian, J., Bath, K., Elison, J., Abbeduto, L., Wolff, J., & Piven, J. (2021). Pre-symptomatic intervention for autism spectrum disorder (ASD): Defining a research agenda. *Journal of Neurodevelopmental Disorders*, 13(1), 49-49.

Haberman, S. J. (1979). *Analysis of Qualitative Data*. vol. 2, new developments. Academic Press.

Hanselman, P., Rozek, C. S., Grigg, J., & Borman, G. D. (2017). New evidence on self-affirmation effects and theorized sources of heterogeneity from large-scale replications. *Journal of Educational Psychology*, 109(3), 405-424.

Harrell, F. (2015). *Regression Modeling Strategies*. Springer.

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: Some applications. *Journal of the American Statistical Association*, 82(398), 371-386.

Hazlett, H. C., Gu, H., McKinstry, R. C., Shaw, D. W. W., Botteron, K. N., Dager, S. R., Styner, M., Vachet, C., Gerig, G., Paterson, S. J., Schultz, R. T., Estes, A. M., Evans, A. C., Piven, J., & the IBIS Network. (2012). Brain volume findings in 6-month-old infants at high familial risk for autism. *The American Journal of Psychiatry*, 169(6), 601-608.

Hazlett, H. C., Gu, H., Munsell, B. C., Kim, S. H., Styner, M., Wolff, J. J., Elison, J. T., Swanson, M. R., Zhu, H., Botteron, K. N., Collins, D. L., Constantino, J. N., Dager, S. R., Estes, A. M., Evans, A. C., Fonov, V. S., Gerig, G., Kostopoulos, P., McKinstry, R. C., . . ., & The IBIS Network. (2017). Early brain development in infants at high risk for autism spectrum disorder. *Nature*, 542(7641), 348-351.

Henderson, C. R. (1963). Selection index and expected genetic advance. *Statistics and Plant Breeding*, 141-163.

Hu, W., Li, G., & Li, N. (2009). A Bayesian approach to joint analysis of longitudinal measurements and competing risks failure time data. *Statistics in Medicine*, 29(11), 1601.

Ibrahim, J. G., Chu, H., & Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 28(16), 2796-2801.

Jacqmin-Gadda, H., Proust-Lima, C., Taylor, J. M. G., & Commenges, D. (2010). Score test for conditional independence between longitudinal outcome and time to event given the classes in the joint latent class model. *Biometrics*, 66(1), 11-19.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1), 35-45.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457-481.

Kikinis R, Pieper SD, Vosburgh K (2014). 3D Slicer: a platform for subject-specific image analysis, visualization, and clinical support. *Intraoperative Imaging Image-Guided Therapy*, Ferenc A. Jolesz, Editor 3(19):277–289

I., E., Kincaid, D., & Cheney, W. (1992). *Numerical analysis–mathematics of scientific computing*. American Mathematical Society.

Kooperberg, C., Stone, C. J., & Truong, Y. K. (1995a). Hazard regression. *Journal of the American Statistical Association*, 90(429), 78-94.

Kooperberg, C., Stone, C. J., & Truong, Y. K. (1995b). The L2 rate of convergence for hazard regression. *Scandinavian Journal of Statistics*, 22(2), 143-157.

Kooperberg, C. (2020). polspline: Polynomial Spline Routines. R package version 1.1.19. [Computer software]. https://CRAN.R-project.org/package=polspline.

Kosorok, M. R. (2009). What's so special about semiparametric methods? *Sankhya. Series. A*, 71(2), 331-353.

Król, A., Mauguen, A., Mazroui, Y., Laurent, A., Michiels, S., & Rondeau, V. (2017). Tutorial in joint modeling and prediction: A statistical software for correlated longitudinal outcomes, recurrent events and a terminal event. *Journal of Statistical Software*, 81(3), 1-52.

Kucharczyk, W., Brant-Zawadzki, M., Sobel, D., Edwards, M. B., Kelly, W. M., Norman, D., & Newton, T. H. (1985). Central nervous system tumors in children: Detection by magnetic resonance imaging. *Radiology*, 155(1), 131-136.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2), 164-168.

Lauterbur, P. C. (1973). Image formation by induced local interactions: Examples employing nuclear magnetic resonance. *Nature*, 242(5394), 190-191.

Letenneur, L., Commenges, D., Dartigues, J. F., & Barberger-Gateau, P. (1994). Incidence of dementia and alzheimer's disease in elderly community residents of south-western France. *International Journal of Epidemiology*, 23(6), 1256-1261.

Liang, K. Y., & Zeger, S. L. (1993). Regression analysis for correlated data. *Annual Review of Public Health*, 14, 43-68.

Lin, H., McCulloch, C. E., & Mayne, S. T. (2002). Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Statistics in Medicine*, 21(16), 2369-2382.

Lin, H., Turnbull, B. W., McCulloch, C. E., & Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, 97(457), 53-65.

Liu, Z., Wang, Y., Gerig, G., Gouttard, S., Tao, R., Fletcher, T., & Styner, M. (2010). Quality control of diffusion weighted images. *Proceedings of SPIE* , 7628(1) 76280J-76280J-9.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129-137.

Logothetis, N. K., Guggenberger, H., Peled, S., & Pauls, J. (1999). Functional imaging of the monkey brain. *Nature Neuroscience*, 2(6), 555-562.

Long, J. D., & Mills, J. A. (2018). Joint modeling of multivariate longitudinal data and survival data in several observational studies of huntington's disease. *BMC Medical Research Methodology*, 18(1), 138-138.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview—Revised: A revised version of a diagnostic interview for caregivers of individuals with possible

pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, 24(5), 659–685.

Losh, M., Martin, G. E., Lee, M., Klusek, J., Sideris, J., Barron, S., & Wassink, T. (2017). Developmental markers of genetic liability to autism in parents: A longitudinal, multi-generational study. *Journal of Autism and Developmental Disorders*, 47, 834-845.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, 10(1), 21-39.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2), 431-441.

Massy, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309), 234-256.

Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology: Vol. 2*. Oxford University Press.

Matthews, P. M., Honey, G. D., & Bullmore, E. T. (2006). Applications of fMRI in translational medicine and clinical practice. *Nature Reviews. Neuroscience*, 7(9), 732-744.

Mersmann, O. (2023). microbenchmark: Accurate Timing Functions v 1.4.10. [Computer software]. https://CRAN.R-project.org/package=microbenchmark.

Miller Jr, R. G. (1981). *Survival analysis*. Wiley.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302), 415-434.

Morrell, C. H., Pearson, J. D., Carter, H. B., & Brant, L. J. (1995). Estimating unknown transition times using a piecewise nonlinear mixed-effects model in men with prostate cancer. *Journal of the American Statistical Association*, 90(429), 45-53.

Mosconi, M. W., Cody-Hazlett, H., Poe, M. D., Gerig, G., Gimpel-Smith, R., & Piven, J. (2009). Longitudinal study of amygdala volume and joint attention in 2- to 4-year-old children with autism. *Archives of General Psychiatry*, 66(5), 509-516.

Munson, J., Dawson, G., Abbott, R., Faja, S., Webb, S. J., Friedman, S. D., Shaw, D., Artru, A., & Dager, S. R. (2006). Amygdalar volume and behavioral development in autism. *Archives of General Psychiatry*, 63(6), 686-693.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1(1), 27-52.

Niu, M., Han, Y., Dy, A. B. C., Du, J., Jin, H., Qin, J., Zhang, J., Li, Q., & Hagerman, R. J. (2017). Autism symptoms in fragile X syndrome. *SAGE Publications*.

Oxelgren, U. W., Myrelid, Å., Annerén, G., Ekstam, B., Göransson, C., Holmbom, A., Isaksson, A., Åberg, M., Gustafsson, J., Fernell, E., Gillberg Neuropsychiatry Centre, Sahlgrenska akademin, Göteborgs universitet, Gothenburg University, Gillbergcentrum, & Sahlgrenska Academy. (2017). Prevalence of autism and attention-deficit–hyperactivity disorder in down syndrome: A population-based study. *Developmental Medicine and Child Neurology*, 59(3), 276-283.

Padilla-Nash, H. M., Heselmeyer-Haddad, K., Wangsa, D., Zhang, H., Ghadimi, B. M., Macville, M., Augustus, M., Schröck, E., Hilgenfeld, E., & Ried, T. (2001). Jumping translocations are common in solid tumor cell lines and result in recurrent fusions of whole chromosome arms. *Genes Chromosomes & Cancer*, 30(4), 349-363.

Papageorgiou, G., Mauff, K., Tomer, A., & Rizopoulos, D. (2019). An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual Review of Statistics and its Application*, 6(1), 223-240.

Pawitan, Y., & Self, S. (1993). Modeling disease marker processes in AIDS. *Journal of the American Statistical Association*, 88(423), 719-726.

Philipps, V., Amieva, H., Andrieu, S., Dufouil, C., Berr, C., Dartigues, J., Jacqmin-Gadda, H., & Proust-Lima, C. (2014). Normalized mini-mental state examination for assessing cognitive change in population-based brain aging studies. *Neuroepidemiology*, 43(1), 15-25.

Piven, J., Nehme, E., Simon, J., Barta, P., Pearlson, G., & Folstein, S. E. (1992). Magnetic resonance imaging in autism: Measurement of the cerebellum, pons, and fourth ventricle. *Biological Psychiatry*, 31(5), 491-504.

Piven, J., Arndt, S., Bailey, J., & Andreasan, N. (1996). Regional brain enlargement in autism: A magnetic resonance imaging study. *Journal of the American Academy of Child and Adolescent Psychiatry*, 35(4), 530-536.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69(2), 331-342.

Prince, M. A., & Fidler, D. J. (2021). Analytic approaches to heterogeneity in neurogenetic syndrome research. In *International Review of Research in Developmental Disabilities* (Vol. 60, pp. 55-73). Academic Press.

Proust-Lima, C., Joly, P., Dartigues, J., & Jacqmin-Gadda, H. (2009). Joint modelling of multivariate longitudinal outcomes and a time-to-event: A nonlinear latent class approach. *Computational Statistics & Data Analysis*, 53(4), 1142-1154.

Proust-Lima, C., & Taylor, J. M. G. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: A joint modeling approach. *Biostatistics* (Oxford, England), 10(3), 535-549.

Proust-Lima, C., Séne, M., Taylor, J. M., & Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research*, 23(1), 74-90.

Proust-Lima, C., Philipps, V., & Liquet, B. (2017). Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *Journal of Statistical Software*, 78(2), 1-56.

Pugliese, C. E., Anthony, L. G., Strang, J. F., Dudley, K., Wallace, G. L., Naiman, D. Q., & Kenworthy, L. (2016). Longitudinal examination of adaptive behavior in autism spectrum disorders: Influence of executive function. *Journal of autism and developmental disorders*, 46, 467-477.

Raschle, N., Zuk, J., Ortiz-Mantilla, S., Sliva, D. D., Franceschi, A., Grant, P. E., Benasich, A.A., & Gaab, N. (2012). Pediatric neuroimaging in early childhood and infancy: challenges and practical guidelines. *Annals of the New York Academy of sciences*, 1252(1), 43-50.

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and Time-to-Event data. *Biometrics*, 67(3), 819-829.

Rizopoulos, D., & Ghosh, P. (2011). A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*, 30(12), 1366-1380.

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

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6(1), 15-32.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.

Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Larsson, H.,Hultman, C. M., & Reichenberg, A. (2014). The familial risk of autism. *Journal of the American Medical Association*, 311(17), 1770–1777.

Schecter, M. T., Craib, K. J. P., Le, T. N., Willoughby, B., Douglas, B., Sestak, P., Montaner, J. S. G., Weaver, M., Elmslie, K. D., & O'Shaughnessy, M. V. (1989). Progression to AIDS and Predictors of AIDS in Seroprevalent and Seroincident Cohorts of Homosexual Men. *AIDS*, 3, 347-353.

Schumann, C. M., Barnes, C. C., Lord, C., & Courchesne, E. (2009). Amygdala enlargement in toddlers with autism related to severity of social and communication impairments. *Biological Psychiatry*, 66(10), 942-949.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461-464.

Shaw, K. A., Maenner, M. J., Baio, J., Washington, A., Christensen, D. L., Wiggins, L. D., Pettygrove, S., Andrews, J. G., White, T., Rosenberg, C. R., Constantino, J. N., Fitzgerald, R. T., Zahorodny, W., Shenouda, J., Daniels, J. L., Salinas, A., Durkin, M. S., Dietz, P. M., & EdS1. (2020). Early identification of autism spectrum disorder among children aged 4 years - early autism and developmental disabilities monitoring network, six sites, United States, 2016. *MMWR. Surveillance Summaries*, 69(3), 1-11.

Shen, M. D., Nordahl, C. W., Young, G. S., Wootton-Gorges, S. L., Lee, A., Liston, S. E., Harrington, K. R., Ozonoff, S., & Amaral, D. G. (2013). Early brain enlargement and elevated extra-axial fluid in infants who develop autism spectrum disorder. *Brain*, 136(9), 2825-2835.

Shen, M. D., Kim, S. H., McKinstry, R. C., Gu, H., Hazlett, H. C., Nordahl, C. W., Emerson, R. W., Shaw, D., Elison, J. T., Swanson, M. R., Fonov, V. S., Gerig, G., Dager, S. R., Botteron, K. N., Paterson, S., Schultz, R. T., Evans, A. C., Estes, A. M., Zwaigenbaum, L., . . ., & Infant Brain Imaging Study Network. (2017). Increased extra-axial cerebrospinal fluid in high-risk infants who later develop autism. *Biological Psychiatry*, 82(3), 186-193.

Signorell, A., et al. (2021). DescTools: Tools for descriptive statistics v 0.99.42. [Computer software]. https://CRAN.R-project.org/package=DescTools.

Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society*, Series B. 13: 238–241.

Stoffer, D., & Poison, N. astsa: Applied Statistical Time Series Analysis. R package version 2.0. [Computer software]. https://CRAN.R-project.org/package=astsa.

Sparks, B. F., Friedman, S. D., Shaw, D. W., Aylward, E. H., Echelard, D., Artru, A. A., Maravilla, K. R., Giedd, J. N., Munson, J., Dawson, G., & Dager, S. R. (2002). Brain structural abnormalities in young children with autism spectrum disorder. *Neurology*, 59(2), 184-192.

Sparrow, S. S., & Cicchetti, D. V. (1989). The Vineland Adaptive Behavior Scales. In C. S. Newmark (Ed.), *Major psychological assessment instruments*, Vol. 2, pp. 199–231). Allyn & Bacon.

Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences*, 22(4), 325-336.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, 5(4), 595-620.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, 22(1), 118-171.

Stone, C. J., Hansen, M. H., Kooperberg, C., & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling: 1994 Wald memorial lecture. *The Annals of statistics*, 25(4), 1371-1470.

Suetens, P. (2009). *Fundamentals of medical imaging*. Cambridge University Press.

Taylor, J. M. G., Yu, M., & Sandler, H. M. (2005). Individualized predictions of disease progression following radiation therapy for prostate cancer. *Journal of Clinical Oncology*, 23(4), 816-825.

Therneau, T. (2021). survival: A Package for Survival Analysis in R. R package version 3.2-11. [Computer software]. https://CRAN.R-project.org/package=survival

Tobii Pro AB (2014). Tobii Pro Lab (Version 1.181) [Computer software]. Danderyd, Sweden: Tobii Pro AB.

Tonnsen, B. L., Boan, A. D., Bradley, C. C., Charles, J., Cohen, A., & Carpenter, L. A. (2016). Prevalence of autism spectrum disorders among children with intellectual disability. *American journal on intellectual and developmental disabilities*, 121(6), 487-500.

Towle, P. O., Patrick, P. A., Ridgard, T., Pham, S., & Marrus, J. (2020). Is earlier better? the relationship between age when starting early intervention and outcomes for children with autism spectrum disorder: A selective review. *Autism Research and Treatment*, 2020, 1-17.

Tsiatis, A. A., De Gruttola, V., & Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, 90, 27-37.

Varadahn, R. & Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scand. J. Statist* 35 335–353.

Veltman, M. W., Craig, E. E., & Bolton, P. F. (2005). Autism spectrum disorders in Prader–Willi and Angelman syndromes: a systematic review. *Psychiatric genetics*, 15(4), 243-254.

Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. *The sage encyclopedia of social sciences research methods*, 2, 549-553.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18(4), 450-469.

Wang, C., Hendricks Brown, C., & Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association*, 100(471), 1054-1076.

Wattjes, M. P. (2011). Structural MRI. *International Psychogeriatrics*, 23(S2), S13-S24.

Williamson, P. R., Kolamunnage-Dona, R., Philipson, P., & Marson, A. G. (2008). Joint modelling of longitudinal and competing risks data. *Statistics in Medicine*, 27(30), 6426-6438.

Wolff, J. J., Gu, H., Gerig, G., Elison, J. T., Styner, M., Gouttard, S., Botteron, K. N., Dager, S. R., Dawson, G., Estes, A. M., Evans, A. C., Hazlett, H. C., Kostopoulos, P., McKinstry, R. C., Paterson, S. J., Schultz, R. T., Zwaigenbaum, L., Piven, J., IBIS Network, & the IBIS Network. (2012). Differences in white matter fiber tract development present from 6 to 24 months in infants with autism. *The American Journal of Psychiatry*, 169(6), 589-600.

Wong, K. Y., Zeng, D., & Lin, D. Y. (2022). Semiparametric latent-class models for multivariate longitudinal and survival data. *The Annals of Statistics*, 50(1), 487-510.

Wulfsohn, M. S., & Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1), 330-339.

Xiao, Z., Qiu, T., Ke, X., Xiao, X., Xiao, T., Liang, F., Zou, B., Huang, H., Fang, H., Chu, K., Zhang, J., & Liu, Y. (2014). Autism spectrum disorder as early neurodevelopmental disorder: Evidence from the brain imaging abnormalities in 2–3 Years old toddlers. *Journal of Autism and Developmental Disorders*, 44(7), 1633-1640.

Xu, J., & Zeger, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics*, 50(3), 375-387.

Zhang, Z., & Sun, J. (2010). Interval censoring. *Statistical Methods in Medical Research*, 19(1), 53-70.