Title
Applications of Shared Random Effects Models to Electronic Health Records Data

Author
Klomhaus, Alexandra Marie-Varnum

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Applications of Shared Random Effects Models to Electronic Health Records Data

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

Alexandra Marie-Varnum Klomhaus

2021

ABSTRACT OF THE DISSERTATION

Applications of Shared Random Effects Models to
Electronic Health Records Data

by

Alexandra Marie-Varnum Klomhaus
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2021
Professor Catherine A. Sugar, Co-Chair
Professor Hilary Aralis, Co-Chair

Electronic health records (EHR) give rise to complicated data structures, due in part to outcome-driven patient and physician decisions that impact the number and spacing of clinical observations, length of time under treatment, and reason for treatment termination. When dependencies exist between patient observations, outcomes, and treatment termination and are ignored in analyses, it can lead to biased parameter estimates and spurious conclusions. Additional complications exist in data obtained from EHR, where informative details, such as reasons for termination of care, frequently go unnoted or remain contained in nonsystematic forms. The objective of this dissertation is to discuss standard difficulties associated with longitudinal analyses using data arising from EHR and to present potential solutions to such challenges using methods appropriate for applied, clinical researchers.

Shared random effects models are a practical and effective method of modeling dependencies between observation times, outcomes, and terminal events. Three-part shared random effects models typically make use of a frailty model for the intensity of observation times of

a medical-related event ("informative observation times"), a general mixed effects model for the longitudinal outcome ("repeated measures") that allows for flexibility in the temporal specification of the overall trajectory, and a Cox proportional hazards model for the timing of termination of care ("dependent terminal event"). This model formulation is typically applied when the dependent terminal event, for example death, is directly observed and distinguishable from independent censoring events. However, termination from care is a terminal event for which the exact date is often unobserved and may be indistinguishable from independent censoring based solely on EHR data, which makes direct application of shared random effects models infeasible.

I propose the use of an inverse cumulative hazard function to estimate individual-level survival times between patients' last-recorded and their next-hypothetical observation, and use these estimates to help classify dependent and independent termination. In a simulation study I illustrate the effectiveness of this method in producing minimally-biased estimates based on the three-part shared random effects model. I apply the same method to model depression symptom trajectories over time using EHR data from Behavioral Health Associates (BHA), a UCLA Health primary and behavioral health collaborative care system.

Further, I examine the utility of a cure model in handling zero-inflated recurrent events data and in an alternate, probabilistic approach to unobserved terminal events we propose an extension of an adaptable cure frailty model that represents the probability that a subject will become unsusceptible to future recurrent events after any given event. I change the terminology from "cure" to "treatment termination" such that I model the probability a patient will terminate treatment after each clinical observation. An added benefit of this approach is the cure model's ability to simultaneously account for the zero-inflatedness common in EHR data (e.g. the overrepresentation of subjects with zero recurrent events).

I describe common issues inherent in EHR data and demonstrate a series of statistical methods that offer practical solutions to these challenges. I provide analytical tools for applied researchers to easily implement such methods in existing statistical software.

The dissertation of Alexandra Marie-Varnum Klomhaus is approved.

Thomas R. Belin

Norweeta G. Milburn

Jessica Jeffrey

Hilary Aralis, Committee Co-Chair

Catherine A. Sugar, Committee Co-Chair

University of California, Los Angeles

2021

To my husband Scott, who has been my constant supporter. As long as I have known you, I have been a student, and I am endlessly grateful for your positivity and your encouragement. Thank you for being the best life partner and dog-parent to Casey. I cannot wait for our next adventure.

To my mom Diana, who was always there to listen as I figured out what I wanted to be when I "grew up". Thank you for instilling in me the importance of independence and education, and for your never-ending love and guidance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The flowchart presented in Figure 2.3 was published in *BMC Health Services Research* [1]. The first author is a member of my committee, Dr. Jessica Jeffrey. I completed the statistical analyses under guidance from co-author and co-chair of my committee, Dr. Hilary Aralis. Work on this paper occurred simultaneously with work on my dissertation.

I would like to acknowledge the patients, staff, clinicians, and faculty associated with Behavioral Health Associates. Access to electronic health records inspired the focus of my dissertation and gave me invaluable knowledge along the way.

I am incredibly grateful for the help and support of my academic advisor and committee co-chair Dr. Catherine Sugar, without whom I may not have decided to pursue a PhD. Her endless faith in my abilities as a statistician and her constant encouragement, both professional and personal, guided me on numerous occasions and is a significant part of how this dissertation ever came to be. I would also like to thank my committee co-chair Dr. Hilary Aralis. We met as students, and watching her developing career gave me inspiration and motivation for my own. I will always be grateful for the time, effort, and guidance she put into helping me through each step of my research and academic career. I would also like to thank Dr. Norweeta Milburn for providing me a space to learn about behavioral health research and teaching me much about setting my own professional goals, Dr. Thomas Belin for being a constant beacon of support and positivity and always looking out for my interests, and Dr. Jessica Jeffrey for helping me understand how to acquaint my statistical abilities with clinical interests. Thank you all for serving on my committee.

To my family, friends, and mentors who have been by my side throughout my academic journey, thank you for the part each of your played in helping me get to this point. To my mom Diana and brother Derek, thank you for your constant support and encouragement. And to my husband Scott, your endless belief in me has meant everything. Thank you for all you do in this life we have built together.

VITA

2012	B.S. (Mathematics and Applied Science) and Minor (Anthropology), UCLA.
2015	M.S. (Biostatistics), UCLA.
2013–2021	Graduate Student Researcher, Department of Psychiatry Division of Population Behavioral Health, Jane and Terry Semel Institute for Neuroscience and Human Behavior, UCLA.
2016–2020	Graduate Student Researcher, Semel Institute Biostatistics Core (SISTAT), Jane and Terry Semel Institute for Neuroscience and Human Behavior, UCLA.
2017	Samuel J. Tibbits Fellowship Recipient, UCLA Fielding School of Public Health.
2015	Teaching Assistant, Department of Biostatistics, UCLA. Biostatistics 200B, under the direction of Professor Thomas Belin.

PUBLICATIONS

Jeffrey J, Klomhaus A, Aralis H, Barrera W, Rosenberg S, Grossman M, Lester P. Factors associated with response and remission from depression at 6-months of treatment in a retrospective cohort treated within an integrated care program. *BMC Health Services Research* 2021; 21(703); https://doi.org/10.1186/s12913-021-06729-1

Milburn N, Klomhaus AM, Comulada WS, Lopez SA, Bath E, Amani B, Jackson J, Lee

A, Kim BKE, Rice E. Reconnecting homeless adolescents and their families: Correlates of participation in a family-based intervention. *Prevention Science*. 2020; 21: 1048–1058; https://doi.org/10.1007/s11121-020-01157-9

Jeffrey J, Klomhaus A, Enenbach M, Lester P, Krishna R. Self-Report Rating Scales to Guide Measurement-Based Care in Child and Adolescent Psychiatry. *Child and Adolescent Psychiatric Clinics of North America* 2020; 29(4): 601–629

Asarnow JR, Tompson MC, Klomhaus AM, Babeva K, Langer DA, Sugar C. Randomized Controlled Trial of Family-Focused Treatment for Child Depression Compared to Individual Psychotherapy: One Year Outcomes. *The Journal of Child Psychology and Psychiatry* 2020; 61(6): 662–671.

Babeva KN, Klomhaus AM, Sugar CA, Fitzpatrick O, Asarnow JR. Adolescent Suicide Attempt Prevention: Predictors of Response to a Cognitive-Behavioral Family & Youth Centered Intervention. *Suicide and Life Threatening Behavior* 2020; 50(1): 56–71

Jeffrey J, Hajal N, Klomhaus A, Marlotte L, Lester P. The Use of Behavioral Health Rating Scales for Primary Care Providers. *Mental Health and Family Medicine* 2019; 15: 884–892.

Bursch B, Lloyd J, Mogil C, Wijesekera K, Miotto K, Wu M, Wilkinson R, Klomhaus A, Iverson A, Lester P. Adaptation and Evaluation of Military Resilience Skills Training for Pediatric Residents. *Journal of Medical Education and Curricular Development* 2017; 4: 1–7.

CHAPTER 1

Introduction

1.1 Electronic Health Records Data

As of 2017, 85.9% of office-based physicians in the United States used some form of electronic health or medical records [2]. As the adoption of electronic health record (EHR) systems becomes more prevalent, we expect that research relying on EHR data will continue to expand. However, EHR systems were originally implemented and intended to support billing and delivery of care, making their more-recent use in research a secondary purpose [3]. This evolution of use means that the format of data from EHR is often incompatible with simple analytic techniques.

EHR data are also structurally complicated, and issues with representativeness have raised questions about its suitability for use in research [4]. Many of these concerns stem from the fundamental differences between EHR data generated from naturalistic observations and longitudinal data that would be collected in a structured research study. In the latter, assessment times are pre-established, often evenly spaced, and chosen based on relevance to the outcomes being measured. Outcomes are collected in a systematic way, typically by individuals blinded to or removed from delivering treatment, and data quality is closely monitored. As such, attempts are usually made to collect information or final outcome measurements from persons who have discontinued treatment. EHR data, particularly when generated in the context of emergent-care or in response to patients' symptoms, may contain redundancies and data collection is often driven by procedural requirements around collec-

tion. The data may be incomplete or tell only a fraction of a patient's story, is often coded for use other than research, and may contain important information in non-systematic forms, such as clinical notes.

As is common in many naturalistic medical settings, outcome-driven patient and physician decisions often impact the number and spacing of patient observations, and hence the timeframe over which observations occur. These decisions include patients' tendencies to seek treatment when feeling unwell and physicians' decisions to schedule and update treatment recommendations based on patients' response to treatment. For example, among patients with chronic pain, those who reported greater pain-related disability were more frequent visitors to primary care [5], while a subsequent review of chronic pain management interventions identified pain severity as associated with treatment dropout [6]. Intensity of health care utilization also varies across follow-up time, evidenced by a greater frequency of visits to health care providers during the disease index year among patients with lupus, and a significant reduction in visits in following years [7].

1.2 Electronic Health Records Data in Behavioral Health

Help-seeking patterns also apply to EHR data in behavioral health interventions, which encompass care and maintenance supporting mental health and resilience, and treatment for substance use disorders [8]. Among patients receiving psychotherapy at a community mental health clinic, those with a depressive disorder were less likely to exhibit treatment attrition and had a higher average number of treatment sessions than those without a depressive disorder, and patients demonstrating long-term symptom improvement were more likely to drop out of behavioral health treatment compared to the baseline termination rate [9]. Additionally, PTSD, depression, and substance use disorders were all individually associated with increased mental health care utilization among veterans enrolled in Veterans Administration primary care [10]. These trends may contribute to the highly irregular observation times

with notable inter- and intra-patient variability present in behavioral health EHR data, contingent upon symptomatology. Indeed, inter-observation times varied dramatically within and across patients receiving care through UCLA's Behavioral Health Associates (BHA), a UCLA Health primary- and behavioral-health collaborative care program (to be described in greater detail in Chapter 2), as depicted in Figure 1.1 for a random subsample of 10 patients.

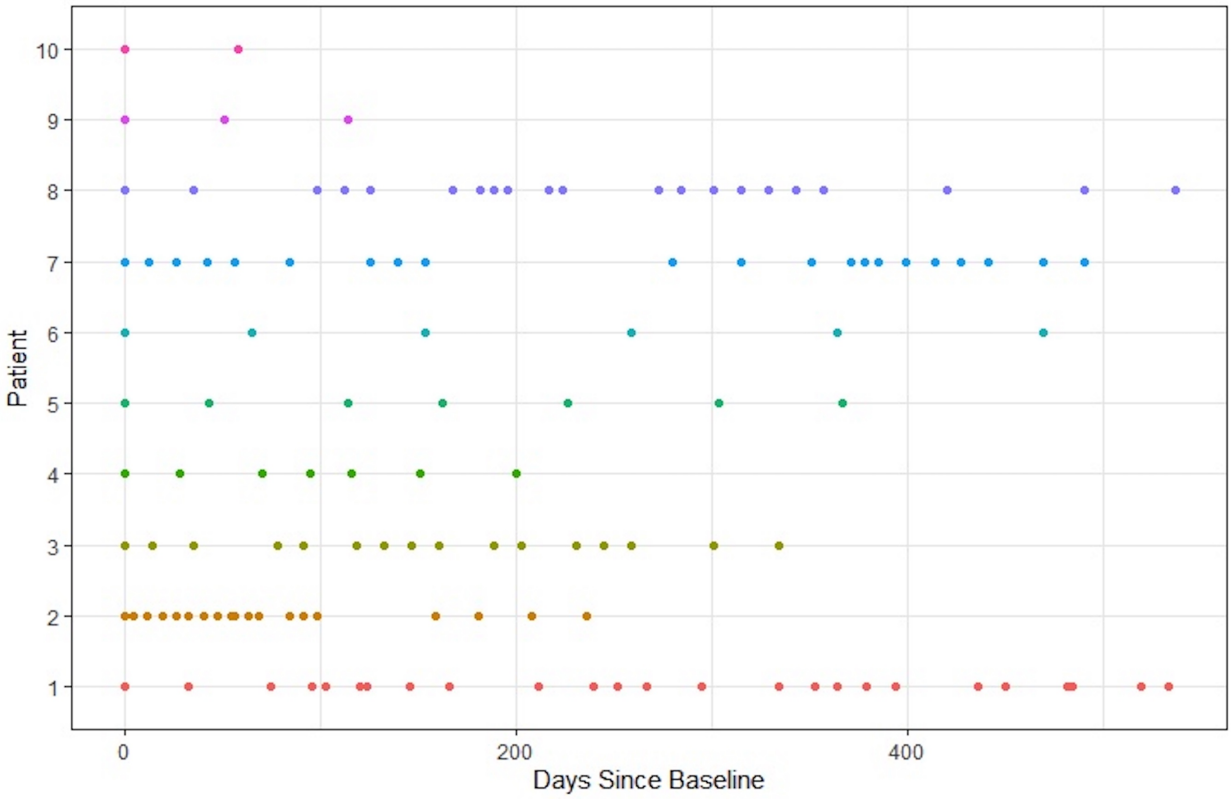


Figure 1.1: Clinical observation times from a random sample of 10 patients under treatment in BHA over their first 548 days of follow-up.

In addition to indicating help-seeking behaviors in response to patients' physical or mental health symptoms, both the intensity and timeframe of clinical observations represent patients' exposure to treatment. In some instances, details of a treatment or therapy might be recorded in the medical record, but in EHR those details may be difficult to access or entirely nonexistent for any given patient record. Behavioral health treatment offers an addi-

tional complication in that therapy sessions do not inherently include a quantifiable measure or dose of treatment [11]. In such instances, clinical observations themselves act as indicators of treatment exposure and their intensity is assumed associated with patients' symptom trajectories insofar as it is assumed that treatment dose impacts a patient's symptoms.

1.3 Informative Observation Times and Dependent Termination

As a result of the help-seeking and outcome-driven care decisions described above, EHR naturally give rise to recurrent event and repeated measures data, often accompanied by a terminal event that precludes further observations [12, 13]. Recurrent events can be generally defined as an ordered set of observations or events, of the same type, at intervals of time for the same individual and include seizure episodes, strokes, and heart attacks, as well as hospitalizations and, as is the case in this paper, clinical observations (e.g. clinical office visits). Repeated measures are defined as quantifiable outcome measurements, again of the same type, observed repeatedly over time and include blood pressure readings, CD4 cell counts, and psychometric measures scores. We often find that the timing of recurrent events, and the presence and timing of a terminal event, are correlated with and carry information about each other in addition to repeated measures, and as such are referred to as "informative observation times" and "dependent termination", respectively [14, 15].

In longitudinal observational data, such as that generated from EHR, we often find that recurrent events, and by association repeated measures, occur at informative observation times. That is, recurrent event times can frequently depend on, or carry information about, the repeated measures themselves, and unobserved confounders or correlations between recurrent event times and repeated measures may remain even after conditioning on covariates [14, 16, 17]. Examples include patients with severe disease symptoms who attend more frequent clinical visits than patients with milder disease symptoms, or physicians recommending more frequent appointments for patients who exhibit worse symptomatology. In

fact, evaluation of EHR data suggests that records with more data points are likely representative of sicker patients [18], resulting in an underrepresentation of healthier patients with more typical observation patterns and moderate repeated measures values.

Simultaneously, it is important to consider that informative observation times and repeated measures may also be accompanied by a terminal event that depends on the observational and/or outcomes processes [13]. When the presence and timing of a terminal event is correlated with recurrent event times or repeated measures, it is often referred to as "dependent termination" and has a non-trivial impact on observations [15]. This terminal event can be directly observed, like death, but as is often the case in medical records data it can also be unobserved but assumed, like treatment termination, withdrawal, or drop-out.

Associations between patient observations, repeated measures, and termination might look like multiple small coronary episodes that ultimately result in a fatal heart attack, or patients attending fewer clinical visits as their symptoms subside, which may positively associate with time-to-termination from treatment. However, the direction of these associations is situation-specific; we can imagine a scenario where patients with severe symptoms attend fewer clinical visits as their disease impedes their ability to travel, ultimately hastening time-to-termination from treatment. In all instances, help-seeking and outcome-related clinical observations in response to emergent medical conditions make standalone longitudinal evaluations of outcomes naïve and prone to bias [13].

Methods of handling these data, particularly when the focus is on understanding longitudinal repeated measure trajectories, depend on properties that relate the observation times, repeated measures, and terminating event processes. In some instances, such as randomized controlled trials where patient observation times are predetermined with little interpersonal variation, simplifying assumptions such as ignoring dependence between recurrent events and repeated measurements are sufficient solutions and traditional analytic techniques may be applied. However, such assumptions are not universally appropriate and may rarely apply when considering EHR data. Even existing methods accounting for terminating processes

rely on observed, recorded terminal event times, but this too is often inapplicable to EHR data where irregularities in patients' observations, including termination of care, can be for a variety of reasons and the timing and/or cause frequently goes unnoted in the medical record.

Shared random effects models have emerged as a practical approach for relating informative observation times, repeated measures, and terminating events. They can be fit using a variety of programmatic approaches and give results that are broadly interpretable. Within this model framework, it is often assumed that each process is independent of the other two, conditional upon the shared random effects [16]. This assumption allows flexibility in the functional specification of each individual process, including the longitudinal time scale. A shared frailty model is commonly used to relate recurrent event times and a terminal event, while the longitudinal component is often fit using a generalized linear mixed model which allows flexible forms for the trajectory, including linear, piecewise, and nonlinear, as well as the specification of within- and between-subject correlation.

An added benefit of shared random effects models is the ability to simultaneously account for another common characteristic of EHR data: zero-inflated recurrent events. Medical records tend to have an overabundance of patients with a singular instance of a health-related event, with no follow-up events or observations. This leads to an excess of "zeros" in the dataset, or those patients with only a "baseline" event and zero "recurrent" events. Shared random effects models can accomodate cure fractions, a logistic approach to describing a sample where some proportion of patients experience some sort of "cure" after their baseline event and are thus unsusceptible to future events.

1.4 Unobserved Treatment Termination

As mentioned above, when a terminal event that precludes further observation in the medical record is associated with observation times and/or repeated outcomes measures, it is

considered a dependent terminal event. The presence and timing of a dependent terminal event can be the result of patient- or physician-driven care decisions, but in many naturalistic observational settings the reason and timing of this terminal event may go unrecorded. For example, studies assessing patient-reported reasons for terminating mental health treatment found the majority of respondants endorsed "feeling better" [19] or perceived ineffectiveness of treatment [20] as the prevailing reasons for termination of care. However, these reasons were reported in a post hoc mental healthcare survey, rather than being reported in the medical record.

While medical records can contain details and timing of observable dependent terminal events, such as death, it is often the case that such information is absent from EHR. Particularly when we consider a terminal event such as treatment termination, it can be the case that the last-recorded patient observation in the medical record is not known to be such. If an outcome-driven decision to terminate treatment was made by either the physician or patient during the visit at the last-recorded observation time, or during the interval between the last and next-hypothetical observation, simply treating the patient as independently censored, or assuming their observations process was inturrupted by an event (e.g. a data pull) entirely unrelated to their observations and/or outcomes processes, leads to inaccurate estimation. Based on patients' individual observation processes and outcome values, we would like to be able to make some assumptions about whether patients dependently terminated treatment or were independently censored.

1.5 Dissertation Overview

This dissertation will explore multiple applications and extensions of shared random effects models for use with EHR data. The first is an application of a three-part shared random effects model to jointly model informative observation times, repeated measures, and dependent termination. Within this model framework, I illustrate a method of classifying

dependent and independent termination when the presence and timing of terminal events is unobserved through use of an inverse cumulative hazard function. I demonstrate the efficacy of this approach through a simulation study, and follow with an application to behavioral health EHR data from UCLA Health's Behavioral Health Associates. I also provide code and guidelines for a SAS macro that implements the described methods while requiring minimal user inputs as a practical tool for applied researchers.

The second application is the use of an adaptive cure frailty model, traditionally used to model preclusion of future recurrent events due to disease "cure", to model the probability of general treatment disconnection when terminal events are unobserved. I also illustrate how a simplified cure model can be used to handle the common occurrence of zero-inflated recurrent events in EHR data in addition to informative observation times, repreated measures, and dependent termination. I present simulation studies to evaluate the efficacy of these applications, and describe a future application to BHA data.

CHAPTER 2

Motivation

2.1 Behavioral Health and Integrated Care in the United States

As of 2017, 46.6 million adults in the US (almost 20%) had some form of mental illness over the past year [21], and approximately 26% of adults had a diagnosable behavioral health condition. Mental health disorders are responsible for 25% of health-related burden and disability worldwide [22], yet regardless of the prevalence many health systems still struggle with effective treatment strategies. Ineffective and underutilized treatment, and poor outcomes, can be attributed to many factors but include systematic segmentation of primary and behavioral health care as well as a general lack of access for many patients in immediate need [23]. A 2009 study found that two-thirds of surveyed primary care providers (PCPs) reported more difficulty in obtaining outpatient mental health treatment for their patients than other commonly utilized services [24], with reasons ranging from patients' health plans to a lack of qualified mental health professionals.

Since the enactment of the Affordable Care Act [25], the US healthcare system has put greater focus on the application and efficacy of integrated care in an attempt to address consequences and complications of a dissociated care structure. The primary goal of many integrated care systems is to coordinate primary and behavioral healthcare, thus providing patients with improved access and treatment for both physical and mental health. In the past, PCPs have often been responsible for providing basic mental health care to their patients, with some PCPs reporting treating as many as 30% of their patients for behavioral

health conditions [26]. Instead, integrated care models view the primary care setting as a pragmatic and often effective screening environment for anxiety, depression, substance abuse, and other mental health disorders with a referral process to a collaborating behavioral health provider for more targeted and timely care. While there are other functional behavioral health treatment models in practice, collaborative behavioral health care is generally more effective in short- and medium-term improvements in primary outcomes, such as depression and anxiety symptomatology, than care from a more segmented system [27].

The increase in integrated care practices suggests that behavioral health observations will more frequently be present in EHR. Though the rate of electronic records adoption in behavioral health lags behind some other specialties like general/family practice and internal medicine, the growth rate of electronic medical records use in psychiatry was 294.2% between 2003 to 2010 [28], a notable trend that should likely continue with increased collaboration and need for information sharing between behavioral health and primary care providers. As such, it is necessary to examine the implications of behavioral health care represented in EHR and specific analytical considerations associated data might require. This includes the fact that behavioral health care can represent a response to an emergent medical condition, rather than routine care like annual physical examinations by a primary care physician, as well as the recognition that treatment can be administered in therapy and thus lack specifically quantifiable dosages. This implies that evaluations of behavioral health treatment programs might incorporate clinical visits as potential indicators of treatment exposure, among other considerations unique to both EHR and behavioral health data.

2.2 Behavioral Health Associates (BHA)

The UCLA Health System, affiliated with the David Geffen School of Medicine at UCLA, is an accountable care organization (ACO), a label given to health provider groups of hospitals, doctors, and other specialists that share responsibility for providing coordinated care

to patients in an effort to reduce unnecessary spending. As UCLA Health's primary care population expands, so does its efforts to effectively address patients' behavioral health concerns. Behavioral Health Associates (BHA), established in 2012, is an effort by UCLA Health to provide wholly integrated primary and mental health care for both adults and children. BHA evolved from a recognized need within the UCLA Health patient population, and itself is a collaborative model alongside UCLA primary care services aimed at providing brief-treatment to patients referred from a primary care provider [29]. BHA includes psychiatry, individual and couples/family psychotherapy, and group psychotherapy in the presence of therapists, psychologists, psychiatrists, social workers, and/or other professionals trained in treating and supporting patients' mental and emotional health and well-being. To support ease-of-access, BHA clinics are co-located in primary care practices.

At the time of BHA's launch in 2012, approximately 21% of patients in UCLA's primary care population had been diagnosed with a behavioral health condition, but UCLA Health was only treating 4% with internal providers [29]. As of the 2016 review of BHA by Clarke et al. [29], BHA had provided behavioral health care to almost 13% of patients with a behavioral health condition and had nearly 190 new referrals, of both adults and children/adolescents, per week.

BHA is intended for short-term behavioral health treatment, with a recommended schedule of one visit every two weeks, up to a total of twelve visits or until six months in care for those patients seeing a therapist. Treatment in BHA should be followed by a referral on to longer-term care or more acute care services, a referral back to primary care for ongoing management, or a resolution of symptoms, depending on a patient's progress at the end of follow-up. Patients' symptoms are assessed via a set of rating scales administered at baseline, and again at approximately three-month intervals. The web-based platform that contains this behavioral health information collected from patients in BHA is called the Behavioral Health Check-up (BHC). Rating scales for adult patients ($\geq$ 18 years) include the Patient Health Questionnaire-9 (PHQ-9) [30] to assess depression, the Generalized Anxiety Disorder-7

(GAD-7) [31] to assess anxiety, the Primary Care PTSD Screen (PC-PTSD) [32] and the PTSD Checklist (PCL) [33] to screen for and assess traumatic stress, and the Alcohol Use Disorders Identification Test (AUDIT) [34] and Drug Abuse Screening Test-10 (DAST-10) [35] to screen for and assess substance use. While clinicians were advised to collect data on these rating scales from patients at least once during each 3-month interval while the patient was receiving treatment, numerous factors impacted the availability of the data including: patient attendance at visits during the desired intervals, workload of the clinical staff required to administer the BHC, and availability of equipment (e.g. tablets) for electronic data collection. However, overall evaluation of outcome measurements and patient information suggests that missing BHC data may be due in large part to administrative challenges and is thus considered to be missing by a mechanism independent of the outcomes process. Additional details of BHA and a more thorough description of the measures used to evaluate patient outcomes have been described elsewhere [1].

2.3 Standard Treatment Evaluations

Assessing the impact of behavioral health treatment programs, like BHA, often includes evaluating change on outcome scores generated from psychometric scales such as those listed above. In BHA, standard reports often include simplified visuals that illustrate the mean outcome score at two time points. This often requires a post hoc decision on which follow-up assessments to use since significant inter-patient variation can exist in the timing of outcome measurements. For example, if we wished to illustrate an improvement in patients' depression symptoms over the course of time under BHA treatment, we might compare an average of patients' baseline PHQ-9 scores to their PHQ-9 scores at the end of follow-up. However, choosing that comparitive PHQ-9 score is challenging as the timing of end-of-treatment varies by patient, and a patient may exit BHA without having been reassessed on the PHQ-9 in many months. A solution might be to simply calculate an average of patients'

last-available PHQ-9 score and compare this to an average of their baseline PHQ-9 score.

Similarly, primary analyses might focus on clinical benchmark improvements, like >50% reduction on the PHQ-9 or a PHQ-9 score < 5, by a certain follow-up assessment. For example, we might be interested in determining how many patients exhibited depression symptom remission (PHQ-9 < 5) at least once in the first six months since their baseline visit to BHA. These analyses can typically be accomplished via simple logistic regression models where the outcome is a binary (yes/no) indicator of patients having reached the threshold of interest. These approaches often assume non-informative observation times and/or independent treatment termination, which simplify the assumed data structure and allow for cross-sectional or simple longitudinal data analysis.

2.4 Challenges and Shortfalls of Standard Treatment Evaluations

These standard evaluations described above often ignore a substantial amount of information that, when incorporated in the analyses, can both reduce bias in resultant parameter estimates as well as illustrate more specifically the totality of the patient treatment experience. For example, a comparison of scores between only two time points ignores the trajectoy of patients' symptoms over the course of follow-up. Likewise, this kind of comparison fails to account for which patients are still in treatment at the time the follow-up score is calculated, and consequently which patients have dropped out of treatment and thus do not contribute to the comparative follow-up score calculation. This is meaningful because, when considering PHQ-9 scores, if patients with lower scores (and thus milder symptoms) or higher scores (and thus more severe symptoms) have a greater tendency to terminate treatment, then the two time point comparison could under- or overrepresent the average change in symptoms over the course of treatment. Likewise, this kind of simplified analysis ignores patients who have only a baseline assessment or who were reassessed on the outcome measure outside the relevant window of time.

Figure 2.1 illustrates mean PHQ-9 scores, among all PHQ-9 scores available, over months since patients' baseline visit, grouped by the total length of time patients were under treatment in BHA. We would like to be able to comment on how the length of time under treatment ultimately affects the overall trajectory of depression symptoms, and by controlling for the general timing of treatment termination we suspect there are differences in depression symptom trajectories based on length of follow-up. For example, we note that among those patients who terminated treatment within the first six months, PHQ-9 scores decreased relatively steadily, with the overall average reaching the clinical benchmark of PHQ-9 < 10 by the third month and ultimately falling below this threshold by the sixth month. We contrast this with patients who terminated treatment between 7-12 months, 13-18 months, and after that year-and-a-half mark, where the average PHQ-9 score reached the clinical benchmark of 10 progressively later (at 4, 5, and 9 months, respectively).

While this kind of subset is helpful in a preliminary visualization of longitudinal data, it ultimately falls short in accuracy and interpretability. We might be concerned about how many PHQ-9 scores contribute to the mean calculations each month or suspect that some of the noise among those who terminated after 18 months is due to the relative sample size. We would like to be able to obtain parameter estimates in a longitudinal analysis that not only account for length of time under treatment but tell us something about the relationship between depression symptoms and risk of treatment termination.

Additionally, outcome measurements may not be collected at every clinical observation, either by design, clinican discretion, or other determination. Figure 2.2 provides a comparison of total patient observations and total outcome measurements collected for a random sample of 10 patients under treatment in BHA. Again, notice the significant inter-patient variation in the timing and frequency of the outcome measurements, and the implications this could have in the ability to comment on the impact of treatment on patients symptoms over the course of follow-up. This figure also illustrates how much information (e.g. the patients' observation times) is ignored in analyses that only focus on the outcome measurements.

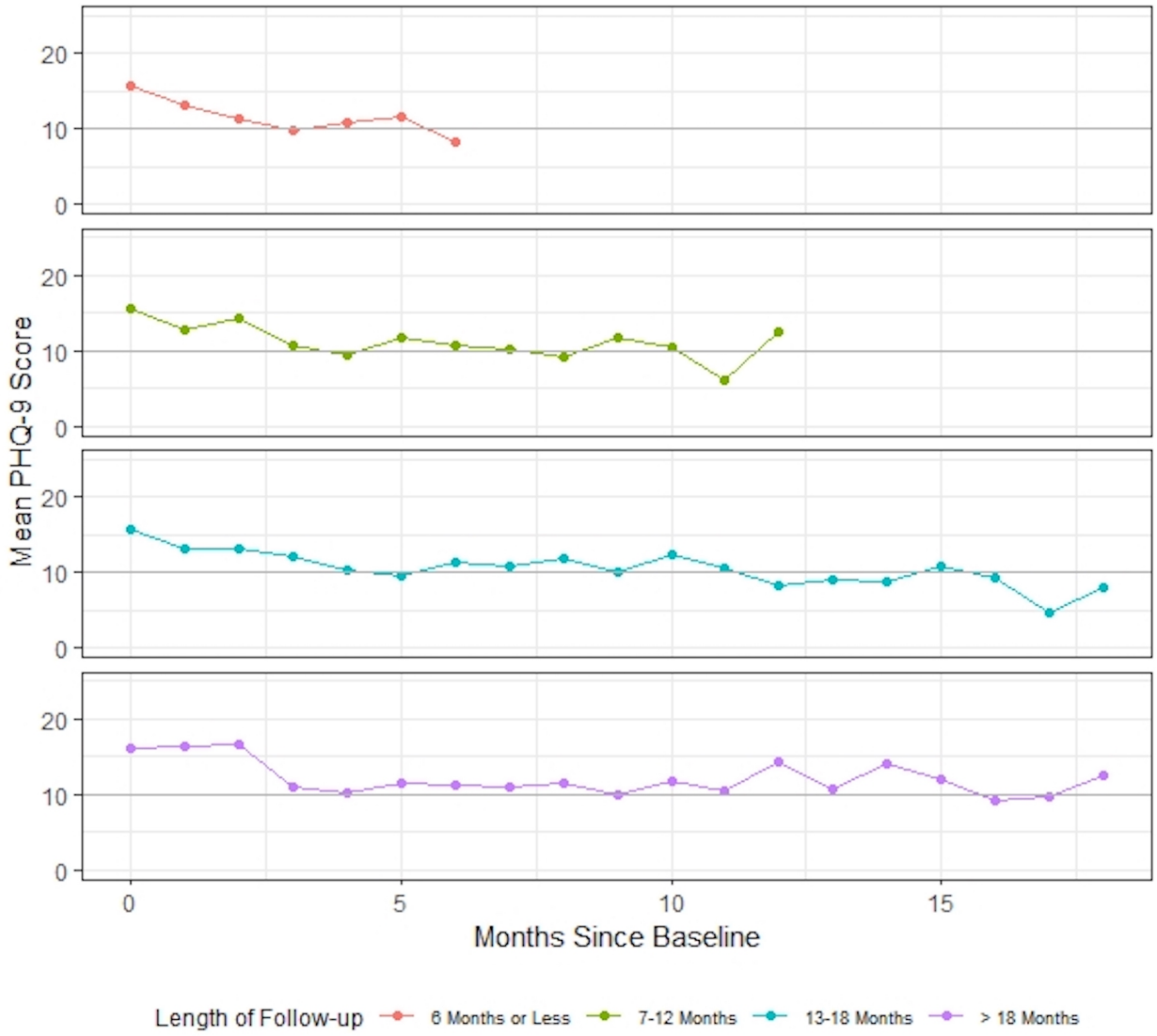


Figure 2.1: Mean PHQ-9 scores among all available at each month of follow-up, subset by total length of follow-up.

Even if the timing of outcome measurements are predetermined, there are numerous reasons why there still may be inconsistencies between patients. Deciding which outcome measurements to use to assess change over time can introduce unforseen complications, and can fail to accurately describe symptom trajectories over the course of follow-up. In the example given in Section 2.3, using patients' last-available PHQ-9 score might mean

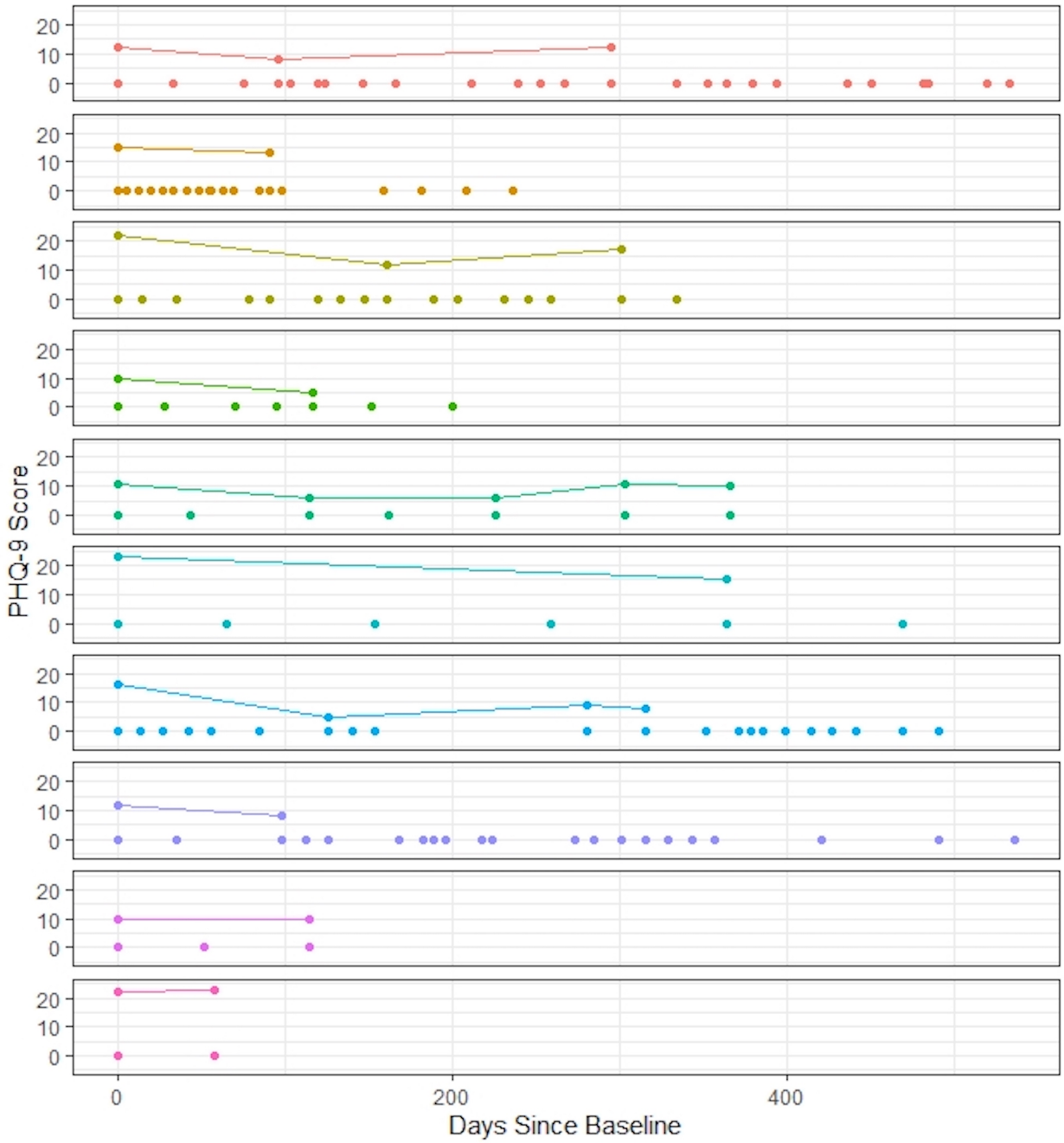
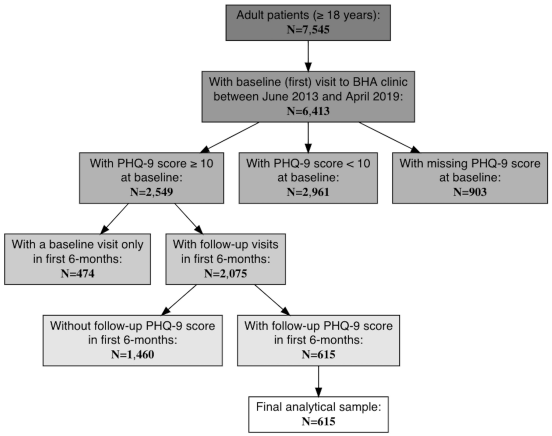


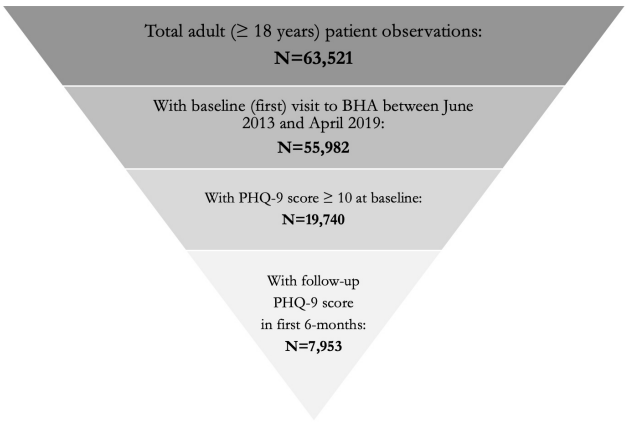
Figure 2.2: Observation times and PHQ-9 scores for a random sample of 10 patients under treatment in BHA.

averaging scores taken at meaningfully different follow-up times (as evidenced in Figure 2.2) and our expectations of changes in depression symptoms vary depending on time spent in treatment.

For a specific example, in a recent paper we used logistic regression to assess treatment response and the presence of depression symptom remission in the first six months of treatment in BHA [1]. The findings suggest important clinical realities, including the impact of comorbid trauma symptoms and suicidal ideation among patients receiving treatment for depression. However, the final analytical sample represents only a portion of patients present in BHA at the time of analysis and uses only a fraction of the total information available for those patients. Figure 2.3 illustrates the construction of the analytical dataset, with the flowchart on the left describing the steps taken to get from the total sample of patients with a baseline visit to the sample of patients with two PHQ-9 scores in the first six months. The chart on the right is a depiction of the corresponding reduction in total available information used, primarily due to ignoring clinical observation times that lacked an outcome measurement. We note that we originally had 63,521 observations corresponding to 7,545 medical record numbers (MRNs), and that we arrived at 7,953 observations corresponding to 615 MRNs once we limited on date, symptomatology at baseline, and the requisite follow-up PHQ-9 score within the first 6 months. Our objective is to include as much of this available information as possible in our analyses, with the ultimate goal of obtaining unbiased estimates that accurately describe patients' symptom trajectories that can be used to describe the totality of patients' treatment experience.



(a) Sample flowchart.



(b) Observation flowchart.

Figure 2.3: Flowcharts for the construction of an analytical sample (used in Jeffrey et al.[1]) evaluating treatment response and depression symptom remission over the first 6-months of treatment in BHA.

CHAPTER 3

Three-Part Shared Random Effects Model for EHR Data

3.1 Background

Preliminary efforts to assess longitudinal covariates while simultaneously evaluating their effect on time-to-event outcomes led to the identification of joint models as a method of accounting for informative censoring and established that considering a covariate over time and concurrently relating it to disease risk reduced parameter estimate bias due in part to informative censoring [36]. Subsequent work continued to demonstrate the potential biases introduced by ignoring dependent terminal events in modeling longitudinal outcomes [37]. Although much of the early literature relating repeated measures and survival assumed noninformative observation times, joint models in the form of shared frailty models were identified as an efficient means of simultaneously considering two correlated survival processes. This allowed for the modeling of associations between recurrent events and terminal events, like in the case of repeated hospitalizations and survival among dialysis patients, or HIV-positive patients [13, 38].

As methods relating recurrent events, repeated measures, and/or censoring developed, researchers began to adapt these methods for implementation in more readily available statistical software, like WinBUGS and SAS. While the use of WinBUGS may offer some programmatic advantage over expectation-maximization (EM) programming [39], it still requires a level of knowledge in Bayesian statistics that is not standard among clinical data

analysts. By estimating unspecified baseline hazard functions using Weibull, exponential, or other parametric functions including piecewise constant functions, resulting fully parametric frailty or shared frailty models can be fit using Gaussian quadrature estimation methods and thus statistical software like SAS and, in particular, SAS's NLMIXED procedure [40]. Estimates obtained through use of Gaussian quadrature methods using parametric model specifications were demonstrated to be comparable to nonparametric estimates using the Monte Carlo EM (MCEM) method, while maintaining the advantage of being more accessible [13]. Whereas other platforms may rely on a depth of algorithmic and Bayesian programming ability, implementation using the SAS NLMIXED procedure (Proc NLMIXED) provides empirical Bayes estimates of random effects without requiring the user to possess more specialized knowledge. Additional benefits of Proc NLMIXED include familiar regression-based parameter estimates and model output, and programmatic requirements that are compatible with applied settings.

Utilizing Gaussian quadrature techniques, Liu et al. (2008) [16] proposed a three-part shared random effects model relating informative observation times, repeated measures, and dependent terminal events, applying the model to hospitalizations and medical cost-accrual data in heart failure patients with disease-related mortality. They echoed previous work and demonstrated via a simulation study that ignoring dependencies between these three processes introduced biases in parameter estimates. Additional literature continued to illustrate the practicality of parametric assumptions, particularly in instances where the longitudinal repeated measures were of primary interest [41]. Alongside the adaptation of methods within the framework of standard statistical software packages, some concurrent and subsequent methods proposed the relaxation of certain distributional assumptions, like the Poisson distributional assumption for recurrent events, and instead advocated for the use of semiparametric specifications for greater flexibility [42–44]. However, the tradeoff for distributional flexibility is again a reliance on MCEM methods, which can unnecessarily complicate implementation and create programmatic and computational barriers not conducive to answering

clinical questions in an applied setting.

3.2 Defining a Three-Part Shared Random Effects Model

Motivated by the objective of developing a model suitable for EHR data in the applied clinical setting, we demonstrate an extension of the three-part shared random effects model proposed by Liu et al. (2008) [16] and apply it to observational EHR data. To differentiate independent terminating events from terminating events that depend on patients' previous observation times and/or their repeated measures, we propose use of an inverse cumulative hazard function estimated from a two-part shared random effects model relating informative observation times and longitudinal repeated measures. We are able to predict whether patients did or did not have sufficient time for another clinical observation prior to the end of our follow-up window based on their individual visit intensities, and am then able to apply the proportional hazards model often used for time-to-death to model time-to-treatment cessation.

3.2.1 Model Notation

Borrowing notation used by Liu et al. (2008) [16], we let T_{ij} denote the j^{th} informative observation time for patient i, where $i = 1, ..., n$ and $j = 1, ..., n_i$. We define $N_{ij}(t) = I(T_{ij} \leq t)$, an indicator function for recurrent events (henceforth referred to as clinical observations), such that $N_i(t) = \Sigma_j N_{ij}(t)$ denotes the total number of clinical observations for patient i occurring at or before time t. Repeated measures, Y_{ij}, are only observed in the presence of a clinical observation, namely when $dN_{ij}(t) = 1$, but are not observed at every observation. Specifically, some proportion of potential repeated measures observations are effectively hidden by a mechanism independent of the longitudinal process and are assumed to be missing completely at random. Thus, we assume that conditional on the occurrence of a clinical observation the probability of observing a repeated measure is independent of

underlying patient outcomes process.

In the case of directly observed independent and dependent terminal events, follow-up time for patient i is stopped at $X_i = min(C_i, D_i)$, the minimum of the independent and dependent termination times, C_i and D_i, respectively. To indicate the presence of a dependent terminal event, we define $\Delta_i = I(D_i \leq C_i)$, where $I(\cdot)$ is an indicator function.

Using this notation, we define a three-part shared random effects model as follows:

1. A frailty model for the clinical observations process, denoted by $r_i(t)$:

$$r_i(t) = r_0(t)exp(\boldsymbol{w_i^R}\beta + u_i) \tag{3.1}$$

2. A mixed effects model for the repeated measures process, denoted by y_{ij}:

$$y_{ij}|(dN_{ij}(t) = 1) = \boldsymbol{z_i}\alpha + \boldsymbol{t_{ij}}\kappa + \gamma_1 u_i + v_i + e_{ij} \tag{3.2}$$

3. A proportional hazards model for the terminal event process, denoted by $\lambda_i(t)$:

$$\lambda_i(t) = \lambda_0(t)exp(\boldsymbol{w_i^C}\eta + \gamma_2 u_i + \gamma_3 v_i) \tag{3.3}$$

where $\{\alpha, \beta, \eta\}$ are unknown parameters and the coefficients associated with the covariate vectors $\boldsymbol{w_i^R}$, $\boldsymbol{z_i}$, and $\boldsymbol{w_i^C}$, for informative observation times, repeated measures, and termination, respectively. Likewise, $\{\kappa\}$ is an unknown parameter and represents the coefficients associated with the longitudinal time vector $\boldsymbol{t_{ij}}$, the contents of which depend on the temporal specification. The baseline hazard function for clinical observations is given by $r_0(t)$, and for dependent termination by $\lambda_0(t)$.

Two random effects, u_i and v_i, are included to account for associations between informative observation times, repeated measures, and dependent termination. Repeated measures depend on clinical observation times via the random effect u_i, while termination depends

on observation times and repeated measures through both u_i and v_i. The unknown parameters $\{\gamma_1, \gamma_2, \gamma_3\}$ are the coefficients on the shared random effects between informative observation times and repeated measures, informative observation times and termination, and repeated measures and termination, respectively. Also of note, u_i and v_i are assumed to be independent of each other, with $u_i \sim^{iid} N(0, \sigma_u^2)$ and $v_i \sim^{iid} N(0, \sigma_v^2)$. We also assume $e_{ij} \sim^{iid} N(0, \sigma_e^2)$, though this model structure can accommodate more complicated within-subject dependencies and correlations between random effects.

Following previous assumptions [16,40], we propose the use of piecewise constant baseline hazard functions to estimate $r_0(t)$ and $\lambda_0(t)$. A piecewise constant baseline hazard function with a sufficient number of nodes has been found to be a satisfactory approximation of the true underlying baseline hazard [45] and the resulting parametric model can be fit using Gaussian quadrature techniques and thus standard statistical software packages like SAS Proc NLMIXED. Nodes can be selected a priori or based on the data itself, and can be defined as evenly spaced intervals, quantiles, or other data-driven timepoints. For specific notation, see Section 3.3. Expected convergence time in Proc NLMIXED depends on model specifications including minimum number of iterations and accuracy of parameters' starting values, as well as size of the dataset.

3.2.2 Likelihood Functions

Specification of the joint likelihood function is of particular importance when implementing a shared random effects model in Proc NLMIXED, as users can specify their own likelihood functions, allowing for additional parametric distributional flexibility.

Under the assumption that the processes described in equations (1)-(3) are independent given the random effects u_i and v_i, again following the notation of Liu et al. (2008) [16] the joint likelihood for the i^{th} patient is:

$$L_i = \int\int l_i^A \, l_i^B \, l_i^C \, p(u_i) \, p(v_i) \, du_i \, dv_i \tag{3.4}$$

where $p(u_i)$ and $p(v_i)$ are the density functions for the random effects u_i and v_i, respectively.

The likelihood contribution for clinical observations for patient i is

$$\begin{aligned} l_i^A = \prod_{j=1}^{n_i} [exp(\boldsymbol{w_i^R}\beta + u_i) r_0(t_{ij})]^{\delta_{ij}} \times \\ exp[-\int_0^{x_i} exp(\boldsymbol{w_i^R}\beta + u_i) r_0(t) dt] \end{aligned} \tag{3.5}$$

where δ_{ij} is an indicator of a clinical observation at time t_{ij}, and x_i is the total observed follow-up time for patient i.

The likelihood contribution for the repeated measures at associated observation times for patient i is

$$l_i^B = \frac{1}{(\sqrt{2\pi}\sigma_e)^{n_i}} \times exp[-\frac{1}{2\sigma_e^2} \sum_{j=1}^{n_i} e_{ij}^2] \tag{3.6}$$

where $e_{ij} = Y_{ij} - \boldsymbol{z_i}\alpha - \boldsymbol{t_{ij}}\kappa - \gamma_1 u_i - v_i$.

The likelihood contribution for the terminal event for patient i is

$$\begin{aligned} l_i^C = [\lambda_0(x_i) exp(\boldsymbol{w_i^C}\eta + \gamma_2 u_i + \gamma_3 v_i)]^{\Delta_i} \times \\ exp[-\int_0^{x_i} exp(\boldsymbol{w_i^C}\eta + \gamma_2 u_i + \gamma_3 v_i) \lambda_0(t) dt] \end{aligned} \tag{3.7}$$

where $\Delta_i = I(D_i \leq C_i)$, an indicator for the presence of a dependent terminal event.

3.3 Piecewise Constant Baseline Hazard Functions

We divide the follow-up time for clinical observations in to M_1 intervals, here defined by every M_1^{th} quantile, and denoted by K_1^R, K_2^R, ...,$K_{M_1}^R$, with $K_0^R = 0$ or the earliest observed event time. We denote the piecewise constant baseline hazard by $\tilde{r}_0(t)$:

$$\tilde{r}_0(t) = \sum_{m=1}^{M_1} r_{0m} I(K_{m-1}^R < t \leq K_m^R) \tag{3.8}$$

with a cumulative baseline hazard of:

$$\tilde{R}_0(t) = \sum_{m=1}^{M_1} r_{0m} max(0, min(K_m^R - K_{m-1}^R, t - K_{m-1}^R)) \tag{3.9}$$

where r_{0m} are a set of M_1 unknown constants to be estimated.

Similarly, we divide the follow-up time for dependent terminal events into M_2 intervals, defined by every M_2^{th} quantile, and denoted by K_1^C, K_2^C, ...,$K_{M_2}^C$, with $K_0^C = 0$ or the smallest observed dependent terminal event time. We denote the piecewise constant baseline hazard by $\tilde{\lambda}_0(t)$:

$$\tilde{\lambda}_0(t) = \sum_{m=1}^{M_2} \lambda_{0m} I(K_{m-1}^C < t \leq K_m^C) \tag{3.10}$$

with a cumulative baseline hazard of:

$$\tilde{\Lambda}_0(t) = \sum_{m=1}^{M_2} \lambda_{0m} max(0, min(K_m^C - K_{m-1}^C, t - K_{m-1}^C)) \tag{3.11}$$

where λ_{0m} are a set of M_2 unknown constants to be estimated.

3.3.1 Gaussian Quadrature

As mentioned previously, the approximation of the baseline hazards using piecewise constant hazard functions results in a parametric joint model that can be fit using Gaussian quadrature techniques and thus Proc NLMIXED in SAS. Other parametric specifications for the baseline hazard functions could be used, such as exponential or Weibull, but we find that piecewise constant functions are a sensible distributional assumption for our eventual application to BHA data.

Moreover, previous research with similar joint models has shown that a piecewise constant baseline hazard function with a suitable number of nodes is a sufficient approximation of the true underlying baseline hazard distribution. In shared random effects models that focus on interpreting the impact of covariates, baseline hazards are something of nuisance parameters so approximations are sufficient in the prioritization of ease of implementation and generalizability.

Although we do not make it an explicit part of our simulation study results in the following section, we will note that we simulated the data under an assumption of a Weibull distribution for the baseline hazard functions for both observation and termination times. We use piecewise constant hazard functions when we implement the shared random effects models and we continue to see sufficiently low biases and high coverage probabilities for our parameter estimates, corroborating the functional use of piecewise constant hazard functions to estimate true underlying distributions in this model formulation.

3.4 Simulation Study Using Inverse Cumulative Hazard to Estimate Dependent and Independent Termination

A distinguishing factor between this and previous work is the lack of precisely observed terminal events, a common occurrence in EHR data. Depending on the nature of treatment, but particularly relevant in instances of treatment in response to emergent medical conditions, we may be able to assume that many patients' observations cease due to a dependent terminal event, like a symptom-based referral or recovery/remission, but the corroborating information often goes unnoted or exists only in non-systematic forms. However, even if they are unobserved, incorporating dependent terminal events in a shared random effects model remains important in accurately evaluating repeated measures trajectories.

We explored the use of a reduced two-part shared random effects model between informative observation times and repeated measures to estimate the time between patients' last-recorded and next-hypothetical clinical observations. We obtained parameter estimates from this two-part model, including population-level estimates for $r_0(t)$ and coefficients for the covariate vector $\boldsymbol{w_i^R}$, and individual-level empirical Bayes estimates for u_i, to calculate this inter-observation time using an inverse cumulative hazard function.

Use of the inverse cumulative hazard function in this way is essentially inverse transform sampling, which is a method for generating a sample from a probability distribution through knowledge of its cumulative distribution function. Using generic notation, to generate survival times using an inversion of the cumulative hazard function we first let X be a random variable with a cumulative hazard fucntion $F_X(x) = 1 - exp(-H(x))$. Then:

1. Generate a random number u from $U \sim Unif(0, 1)$.
2. Find $F_X^{-1}(x)$.
   (a) Using the cumulative hazard above, this yields $F_X^{-1}(x) = H^{-1}(-log(1 - x))$.
3. Calculate $X = F_X^{-1}(u)$.

(a) Using the equation in (a), this gives $X = H^{-1}(-log(1-u))$.

Because $-log(1-U) \sim Exp(1)$, note that it is possible to apply the inverse cumulative hazard function to an $Exp(1)$ random variable.

Like the primary three-part shared random effects model, we made the distributional assumption in this reduced two-part model formulation that the baseline hazard function for observation times would be piecewise constant. Again using generic notation, inverting a piecewise constant cumulative hazard function with two knots might look like:

$$h(t) = \begin{cases} f_1, 0 \leq t \leq t_1 \\ f_2, t_1 < t \leq t_2 \\ f_3, t > t_2 \end{cases}$$

with a corresponding cumulative hazard function of:

$$H(t) = \begin{cases} f_1 t, & 0 \leq t \leq t_1 \\ f_1 t_1 + f_2(t - t_1), & t_1 < t \leq t_2 \\ f_1 t_1 + f_2(t_2 - t_1) + f_3(t - t_2), & t > t_2 \end{cases}$$

An inversion of this cumulative hazard function gives:

$$H^{-1}(x) = \begin{cases} \frac{x}{f_1}, & 0 \leq x \leq f_1 t_1 \\ t_1 + \frac{x - f_1 t_1}{f_2}, & f_1 t_1 < x \leq f_1 t_1 + f_2(t_2 - t_1) \\ t_2 + \frac{x - f_1 t_1 - f_2(t_2 - t_1)}{f_3}, & x > f_1 t_1 + f_2(t_2 - t_1) \end{cases}$$

where the f's represent a generic functional form. An extension with additional knots follows the same form. Evaluation of this inverted cumulative hazard function yields survival times we can use to estimate the time between patients' last-recorded observation and their next-hypothetical observation.

After predicting a next-hypothetical visit time for each patient, relative to their baseline visit time, we compared this to a predetermined independent censoring time (e.g. a data extraction time) to distinguish between patients who had sufficient time to return for another clinical observation but did not, and were hence considered dependently terminated, and patients who did not have sufficient time to return between their last-recorded observation and the independent censoring time, who were hence considered independently terminated. Assuming T_{in_i} represents patient i's last-recorded observation time and $T_{i(n_i+1)}$ their next hypothetical observation time, we update the indicator function for a dependent terminal event (defined in Section 3.1) such that $\hat{\Delta}_i = I(T_{i(n_i+1)} \leq C_i)$ and follow-up time such that $\hat{X}_i = \hat{\Delta}_i \times D_i + (1 - \hat{\Delta}_i) \times C_i$. Users might make small adjustments to the C_i boundary to adjust for certain dataset characteristics as well as some sampling uncertainty introduced using this method. We will refer to this as the "inverse cumulative hazard method".

To demonstrate the accuracy of this described method, we simulated G=200 dataset replicates and predicted dependent and independent termination, then ran the proposed three-part shared random effects model. We calculated average parameter estimates, as well as percentage biases and coverage probabilities across all replicates. We compared these results to those obtained using known terminal event types, as well as with results obtained by treating all terminal events as independent, effectively setting $\gamma_2 = \gamma_3 = 0$.

As a naïve contrast, we predicted termination type by comparing each individual's longest interval between two consecutive observations to the amount of time between their last-recorded observation and the independent termination time. Those whose longest time between two observations exceeded the time between their last observation and the independent termination time were considered independently terminated, while all others were

considered dependently terminated. We will refer to this as the "longest visit gap method". As a reference, original notation changes to accomodate the longest visit gap method in the following ways. Continuing to assume T_{in_i} represents patient i's last-recorded observation time and C_i the independent censoring time, the indicator for a dependent terminal event is now $\hat{\Delta}_i$=$I(max_i(T_{i2} - T_{i1}, ..., T_{ij} - T_{ij-1}, ..., T_{in_i} - T_{in_i-1}) \leq C_i - T_{in_i})$. With this update to the notation for $\hat{\Delta}_i$, we continue to determine patient i's follow-up time by $\hat{X}_i$=$\hat{\Delta}_i \times D_i + (1 - \hat{\Delta}_i) \times C_i$. Although we will focus on the method using the inverse cumulative hazard function, we wished to provide the corresponding notation for the longest visit gap method as there are scenarios in which this simpler approach may be sufficient or even preferrable.

Referencing equations (1)-(3) in Section 3.1, we define z_i, w_i^R, and w_i^C to be the same fixed binary covariate at the individual level that can take values of 0 or 1 each with a probability of 0.5. We set β=1, η=1, and $\alpha = (10, 1)^T$, while we set $\kappa = 0.2$. We set the error term $e_{ij} \sim N(0, \sigma_e^2)$ with $\sigma_e^2 = 1$. For each individual, a baseline repeated measure is defined by y_{i0} at time t_{i0}, and a repeated measure is observed at each subsequent observation time. Repeated measures are not observed at either independent or dependent terminal event times. The random effects $u_i \sim^{iid} N(0, \sigma_u^2)$ and $v_i \sim^{iid} N(0, \sigma_v^2)$ with $\sigma_u^2 = 1$ and $\sigma_v^2 = 0.5$, and further assumed $(\gamma_1, \gamma_2, \gamma_3)^T = (1.5, -0.5, 0.5)^T$.

We simulated both informative observation and dependent termination times under Weibull distributional assumptions. For the baseline hazard for informative observation times, we set the shape parameter to 1 and the scale parameter to 0.01, and for the baseline hazard for the dependent terminal event we set the shape parameter to 2.25 and the scale parameter to 0.00265. Lastly, we used 360 + Uniform(0,6) to simulate dataset-wide independent termination times. These parameters were chosen to produce data on the same general scale as the observed data in the following application in Chapter 4. The estimation method in Proc NLMIXED assumes piecewise constant baseline hazard functions, but this and prior work demonstrates the functionality of piecewise constant baseline hazard

functions as an approximation of the true underlying distribution. We used 5 quantiles to define the nodes of the baseline hazard function for informative observation times in the 2-part model used to obtain estimates for the inverse cumulative hazard, and 10 quantiles to define the nodes of the baseline hazard functions for both the informative observation and dependent termination times in the full 3-part model.

Across the 200 dataset replicates, use of both termination prediction methods produced similar results in terms of classification accuracy, with an average of 85% of termination types correctly classified using either method. A notable difference lies in which direction each method tends to misclassify. Among individuals for whom the method of prediction misclassified their termination type, the inverse cumulative hazard method favored classifying individuals as dependently terminated when they were truly independently terminated, and the longest visit gap favored classifying individuals as independently terminated when they were truly dependently terminated. This particular finding may partially depend on defined parameter values, which were chosen to loosely coincide with the observed data in our following application. The average classification rates using the inverse cumulative hazard method are depicted in Figure 3.1.

Results from the shared random effects models across the 200 dataset replicates using the four terminal event scenarios are found in Table 3.1. Predicting termination type using the inverse cumulative hazard method resulted in a percentage bias of 10.0% and a coverage probability of 87.5% for γ_2. This finding is not unreasonable, as the proposed method is an estimation of terminal events that relies on results from a two-part model between informative observation times and repeated measures, which was already argued to be potentially biasing when used in place of a three-part model. However, the tradeoff between a marginal bias in γ_2 and being able to incorporate the necessary information to use a proportional hazards model for terminal events, and thus a three-part shared random effects model, favors the latter. Bias is largely limited to γ_2, and coverage probabilities of the remaining parameters, including the covariate effects, are all larger than 90%. Predicting termination type using the longest visit

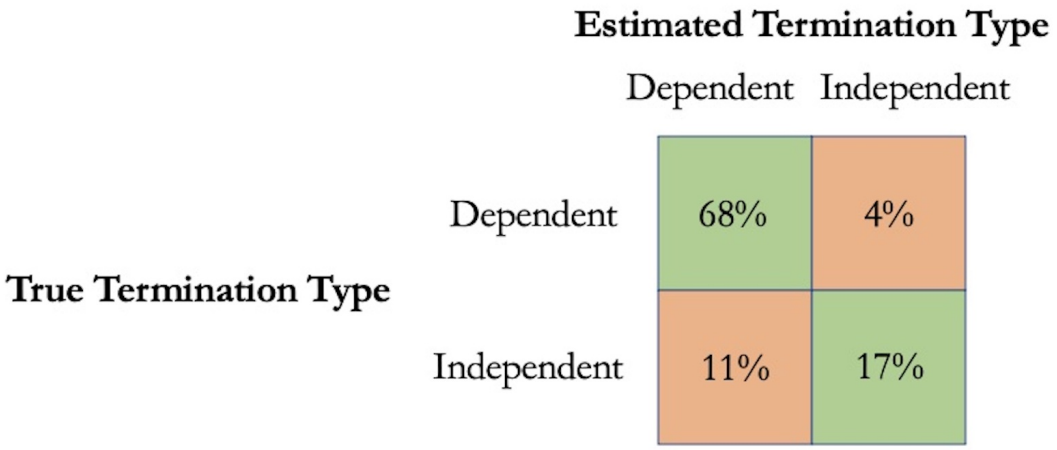


Figure 3.1: The average termination classification rates using the inverse cumulative hazard method across the 200 dataset replicates. Correct classification is denoted in green, and incorrect classification is denoted in red.

gap method did result in additional, and somewhat larger, percentage biases, including 22.4% for γ_2, with a coverage probability of 85.5%, and 22.3% for γ_3, with a coverage probability of 86.0%. The percentage biases on the remaining parameters were also relatively small, with the largest being 7.1% for η, although the accompanying coverage probability is 91.5%. Treating all terminal events as independent resulted in the lowest coverage probabilities for the associated parameters, including 78.0% for γ_1, 81.5% for σ_u^2, 78.5% for σ_v^2, and 89.0% for κ. Although not included here, ignoring terminal events altogether resulted in even greater bias and reduced coverage probability for the associated parameters.

Predicting termination using the inverse cumulative hazard method is notably more precise, in terms of parameter estimates with minimal percentage bias and maximal coverage probability, than treating all terminal events as independent, with performance approaching that of the oracle scenario. It was also somewhat better than the alternative prediction scenario presented, though even this more naïve method continued to be preferrable to treating all terminal events as independent. There may be situaitons in which even a simple method of independent/dependent termination discrimination, like the longest visit gap method pre-

sented here, is a sufficient solution to unobserved terminal events. We present our application using the inverse cumulative hazard method.

We wrote a SAS macro that can be used to implement the described model, and provide a description and sample implementation in Chapter 7 and the macro code in the Appendix.

Table 3.1: Simulation results from implementation of a three-part shared random effects model using different approaches to handling unobserved terminal events.

	True Termination			**Inverse Cumulative Hazard[1]**			**Longest Visit Gap[2]**			**Reduced 2-Part Model[3]**		
Parameter	**Estimate[4]**	**%Bias[5]**	**CP[6]**	**Estimate**	**%Bias**	**CP**	**Estimate**	**%Bias**	**CP**	**Estimate**	**%Bias**	**CP**
						Informative Observation Times						
$\beta = 1$	0.988	1.20	95.00	0.989	1.10	95.50	0.957	2.50	95.00	1.003	0.30	90.50
						Repeated Measures						
$\alpha_0 = 10$	10.012	0.12	94.00	10.011	0.11	94.50	10.031	0.02	94.00	10.015	0.15	90.00
$\alpha_1 = 1$	0.978	2.20	95.00	0.979	2.10	95.50	0.932	4.60	95.00	0.972	2.80	91.00
$\kappa = 0.2$	0.200	0.00	93.00	0.200	0.00	93.50	0.200	0.00	93.50	0.200	0.00	89.00
$\sigma_e^2 = 1$	1.001	0.10	95.00	1.001	0.10	95.50	0.999	0.20	95.00	1.002	0.20	91.50
						Termination						
$\eta = 1$	1.023	2.30	94.00	1.005	0.50	94.50	0.929	7.10	91.50			
						Model Association						
$\gamma_1 = 1.5$	1.513	0.87	97.00	1.505	0.33	98.00	1.478	1.47	96.50	1.571	4.73	78.00
$\gamma_2 = -0.5$	-0.494	1.20	93.00	-0.450	10.00	87.50	-0.388	22.40	85.50			
$\gamma_3 = 1$	1.008	0.80	95.00	0.937	6.30	92.00	0.777	22.30	86.00			
						Variance/Covariance						
$\sigma_u^2 = 1$	0.979	2.10	90.50	0.986	1.40	91.50	1.008	0.80	90.00	0.924	7.60	81.50
$\sigma_v^2 = 0.5$	0.480	4.00	90.50	0.488	2.40	92.00	0.488	2.40	93.50	0.438	12.40	78.50

[1] Results from 3-part model where independent or dependent termination predicted using the inverse cumulative hazard method.

[2] Results from 3-part model where independent or dependent termination predicted using the longest visit gap method.

[3] Results from reduced two-part model treating all terminal events as independent; equivalent to constraining γ_2 and γ_3 equal to 0.

[4] Average parameter estimates across G=200 dataset replicates.

[5] Calculated as: abs((true parameter value - estimated parameter value) / true parameter value) x 100

[6] CP: coverage probability; calculated as: the percentage of 95% confidence intervals that contained the true parameter value.

CHAPTER 4

Application of Three-Part Shared Random Effects Model to BHA Data

4.1 BHA Sample Description

To demonstrate performance of this shared random effects model on EHR data with prediction of independent and dependent terminal events, we apply it to data obtained from Behavioral Health Associates (BHA), a collaborative primary care and behavioral health program part of the UCLA Health System. Patients enter BHA following a referral from their UCLA Health primary care provider (PCP). Data include patients who initiated a new treatment episode and had an associated baseline visit date between March 2013 and October 2019. We included all visits corresponding to each patient's first new treatment episode and excluded visits labeled for subsequent episodes. Patient data were obtained from two sources: CareConnect, UCLA's EHR system, and the Behavioral Health Check-up (BHC), a web-based platform integrated within the EHR which consists of electronic behavioral health data collected from UCLA patients who attended at least one visit to a BHA clinic. Information from CareConnect and the BHC were linked via patients' medical record numbers (MRNs), a unique patient identifier, and only patients with a record in both databases were included in the analytical sample.

The sample was limited to adult patients ≥ 18 years of age at the time of their first visit to BHA for a new treatment episode after March 2013. We included only those patients that had been seen at BHA with a visit labeled as "New", which became their baseline visit,

and excluded all patient encounters not listed as "completed" in the data. We aimed to restrict the sample to a single treatment episode for each patient consisting of a patient's baseline visit and all follow-up visits that were a result of the same referral and for the same behavioral health condition. To do so, we excluded any subsequent visits labeled as "New" occurring more than 180 days after the patient's previously-recorded visit to BHA. Since this paper focuses on depression as the primary outcome, the sample includes those patients with a PHQ-9 score ≥ 10 at baseline as a proxy for those receiving some degree of depression treatment in BHA.

To allow for longitudinal evaluation of symptoms in all patients, the analytical sample was further restricted to those with a baseline PHQ-9 score and at least one additional PHQ-9 score during follow-up. This resulted in exclusion of a substantial number of patients with only a baseline PHQ-9 score whom we assumed to be systematically different from the target population of patients receiving treatment in BHA (this point is further addressed in the discussion). It is also important to note that that our repeated measure, patients' PHQ-9 scores, is not observed at every clinical observation. Specifically, some proportion of our potential repeated measures are effectively hidden by a mechanism independent of the longitudinal process and are assumed missing by design based on the considerations noted in Section 5.1. Thus, we assume that conditional on the occurrence of a clinical observation the probability of observing a repeated measure is independent of the underlying outcomes process.

Lastly, because our primary clinical interest is in patients' depression symptom trajectories under treatment in BHA, we restricted our follow-up time to 1.5 years (or 548 days) following a patient's baseline visit. This decision addresses concerns regarding sparsity in both clinical observations and repeated measures over timeframes extended beyond 1.5 years and focuses our analyses on short- and moderate-term treatment effects more accurately reflecting the clinical intention of the BHA treatment program.

The resulting analytic dataset contained 949 patients and 10,590 observations with a

median of 9.0 (IQR = 8.0) visits to BHA during the first year-and-a-half of treatment. The median follow-up time from baseline to a terminal event was 339.0 days (IQR = 367.0). Summary statistics of BHA visit information can be found in Table 4.1.

Table 4.1: Summary statistics of model-related components across patients' first 548 days under BHA treatment.

	Mean (SD)[1]	**Median**	**IQR**[2]
First-Available PHQ-9 Score	15.85 (4.53)	15.00	7.00
Last-Available PHQ-9 Score	10.48 (6.39)	10.00	10.00
Visits to BHA (Overall)[3]	11.16 (8.03)	9.00	8.00
Visits to BHA Physician	4.92 (3.90)	4.00	6.00
Visits to BHA Therapist	5.54 (7.61)	2.00	9.00
Time in Treatment	344.63 (176.14)	339.00	367.00

[1] SD: standard deviation
[2] IQR: interquartile range
[3] Some BHA visits are missing a provier-type label

4.2 Model Specification for BHA Data

In the longitudinal model, we control for demographic variables age, marital status, gender, and race/ethnicity, where the average age of our sample is 41.2 years (sd = 14.8) and 34.4% are married. Our sample is also majority female (67.5%) and white/Caucasian (57.9%). Previous research suggests that baseline trauma symptoms and baseline suicidal ideation are significantly associated with decreased odds of depression symptom remission over the first 6-months of depression treatment in BHA [1], so we include baseline behavioral health condition indicators for suicidal ideation (positive endorsement on PHQ-9 Item 9) and trauma symptoms (PCL > 50). Other covariates include an indicator for baseline anxiety (GAD-7 ≥ 10), as well as an indicator for whether patients saw a physician, compared to a therapist, for the majority ($> 50\%$) of their visits to BHA.

Within our sample, 72.2% of patients had comorbid anxiety at baseline, 40.4% had some suicidal ideation in the weeks prior to their baseline visit, and 28.8% had comorbid trauma

symptoms, while 50.6% saw a physician for the majority of their visits. To test the hypothesis of differential depression treatment response by comorbid trauma symptoms [1], we also included interactions between trauma symptoms and our time variables to allow for potentially divergent depression symptom trajectories. We include the indicator for whether patients saw a physician for $> 50\%$ of their visits as a covariate in both the informative observation times and dependent treatment termination models, in addition to the longitudinal outcomes model.

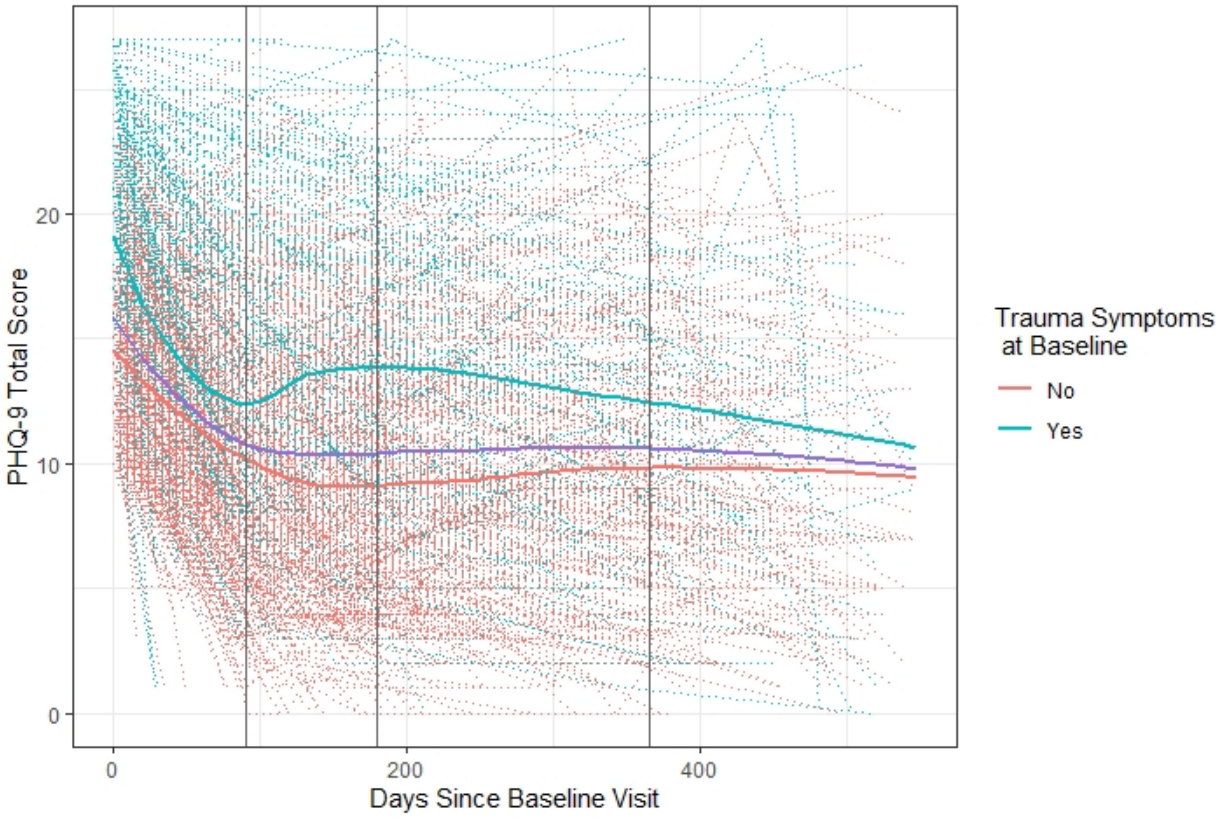


Figure 4.1: Raw depression symptom trajectories among patients under treatment in BHA. LOESS curves are shown for patients with and without comorbid trauma symptoms at baseline, and overall (purple curve). Vertical bars represent 90, 180, and 365 days since baseline.

Upon consideration of the BHA treatment and assessment schedule, and examination of unadjusted depression symptom trajectories over follow-up, we ultimately determined that

a piecewise linear temporal specification for the time variable within the repeated measures mixed effects model was most appropriate. As mentioned previously, the rating scales are meant to be administered at baseline and again every three months, meaning that the presence of repeated measures should roughly correlate with 90 and 180 days post-baseline. Our data also support the clinical experience that many patients stay in treatment longer than the intended six months, so we added an additional knot at 365 days, with the increased spacing reflecting longer-term stay but with some decrease in visit and repeated measure intensities. Figure 4.1 provides a visual depiction of where these knots fall relative to depression symptom trajectories over the course of follow-up, with the green and red loess curves representing the trends among patients with and without comorbid baseline trauma symptoms and the purple curve the overall trend. We used 5 quantiles to define the nodes of the baseline hazard functions for both the informative observation and dependent terminal event times based on inspection of both the intensity of clinical observations and dependent terminal event times over the follow-up window.

We use the inverse cumulative hazard, calculated from estimates obtained by fitting a two-part shared random effects model between informative observation times and repeated measures, to predict dependent and independent termination in our sample. Restricting our follow-up time to 1.5 years, we define $C_i' = min(549, C_i)$ such that $\hat{\Delta}_i = I(T_{i(n_i+1)} < C_i')$. Since exact termination time is not recorded, we use knowledge of the BHA program to assign a termination time of one day after patients' last recorded visit to BHA. We consider this assumption equivalent to decisions made during a patient's last visit, whether it be in response to their symptomatology or other clinical observations, that resulted in some form of termination of care in BHA. This suggests we update the definition of our follow-up time to $\hat{X}_i = \hat{\Delta}_i \times (T_{in_i} + 1) + (1 - \hat{\Delta}_i) \times C_i'$.

4.3 Results of Application to BHA Data

The parameter estimates from this application are shown in Table 4.2, while contrasts evaluating depression symptom trajectories can be found in Table 4.3. Overall, time under treatment in BHA is significantly associated with a decrease in depression symptoms, relative to baseline, regardless of comorbid trauma status with significant improvement evident in trajectories across both 1-year (p=0.003) and 1.5-year (p=0.002) follow-up windows. Patients with and without comorbid trauma symptoms had a significant improvement in depression symptoms in the first 3 months of treatment, with patients with comorbid trauma symptoms entering BHA with more severe depression ($\alpha_5 = 2.935$, $p < 0.001$) but also demonstrating significantly greater improvement than those without comorbid trauma symptoms ($p < 0.001$). Both patient groups had a significant change in their trajectories at the 3-month changepoint (ps < 0.001) and a subsequent divergence in trajectories between 3 and 6 months ($p = 0.005$). Patients without comorbid trauma symptoms demonstrated a significant leveling-off and no significant improvement between 3-6 months, while patients with comorbid trauma symptoms displayed a temporary worsening of depression symptoms ($est. = 0.020$, p=0.022). Patients with comorbid trauma symptoms had another significant change in trajectory at the 6-month mark (p=0.020) and consequently maintained a gradual improvement in depression symptoms for the remainder of follow-up.

Figure 4.2 illustrates the described temporal trends using estimated fixed effects means, evaluated at sample averages listed in Section 4.2. We also note that by the 3-month follow-up mark, the estimated PHQ-9 means regularly hover around 10, often considered a clinical threshold for depression treatment "response", among patients without comorbid trauma symptoms. However, it takes roughly 1.5 years for the estimated means among patients with comorbid trauma symptoms to reach a similar threshold.

Comorbid baseline behavioral health conditions and other patient characteristics were also significantly associated with patients' depression symptom trajectories. In addition to

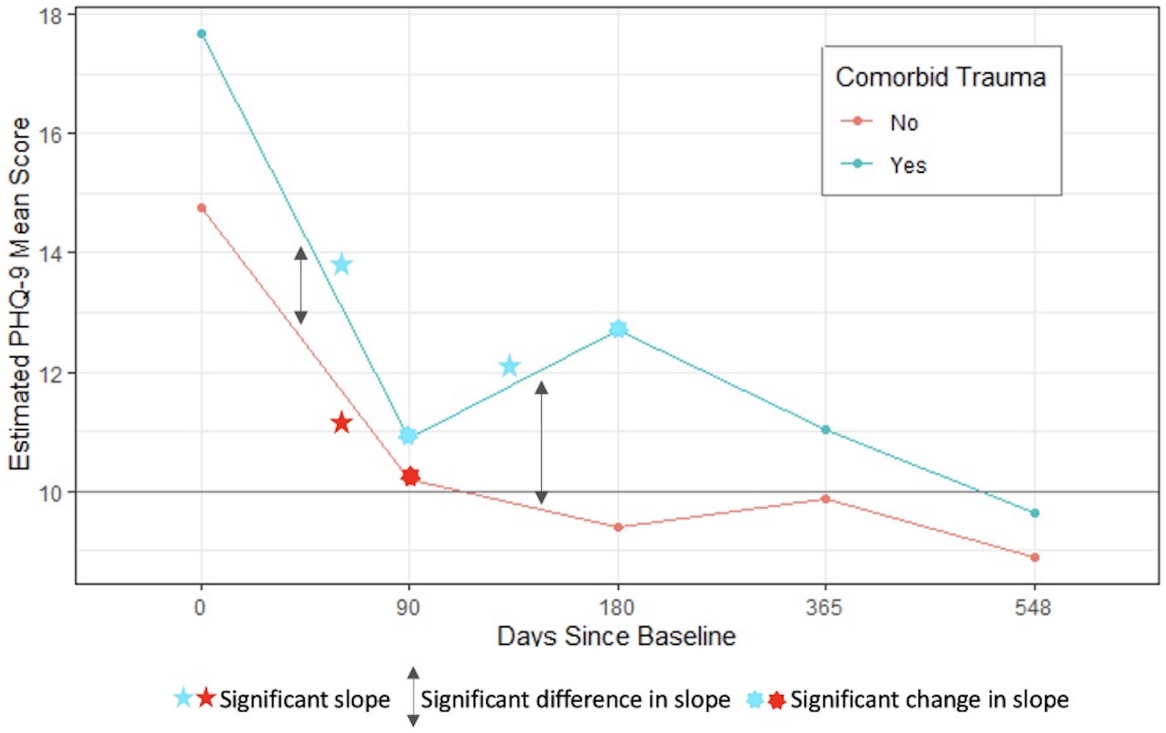


Figure 4.2: Estimated PHQ-9 means at 0, 90, 180, 365 and 548 days, evaluated using sample averages of covariate values. Horizontal line at 10 represents clinical threshold for treatment "response", or "mild" depression symptoms.

trauma, the presence of comorbid baseline anxiety ($\alpha_6 = 2.160$, $p < 0.001$) and suicidal ideation ($\alpha_7 = 2.440$, $p < 0.001$) were both associated with higher PHQ-9 scores and thus more severe depression symptoms. We did not find race/ethnicity, age, gender, or marital status to be significantly associated with PHQ-9 scores, however this could be due in part to the demographic homogeneity of our sample. We did find majority provider (physician vs. therapist) to be significantly associated with all three processes. Patients who saw a physician for the majority of their visits to BHA had significantly lower intensity of visits (β = -0.904, $p < 0.001$), more severe depression symptoms ($\alpha_8 = 0.564$, p=0.030), and decreased risk of dependent treatment termination (η = -0.569, $p < 0.001$) compared to those who saw a therapist for the majority of their visits.

Further, we note that the coefficients on the shared random effects are all significant,

supporting our use of this joint model. Estimates suggest that observation times are indeed informative, and that patients with a greater intensity of visits to BHA tended to have more severe depression symptoms ($\gamma_1 = 2.184$, $p < 0.001$). Termination depended on both the observation times and PHQ-9 scores, with results suggesting that patients with a greater intensity of visits to BHA ($\gamma_2 = 0.591$, $p < 0.001$) and/or patients with less-severe depression symptoms (γ_3 = -0.058, p=0.007) had a greater risk of dependent treatment termination. Lastly, the variance components for each of the random effects suggests heterogeneity for both the patient observation times ($\sigma_u^2 = 0.101$, $p < 0.001$) and the repeated PHQ-9 scores ($\sigma_v^2 = 7.473$, $p < 0.001$).

For model checking, we used the 'predict' statement in Proc NLMIXED in calculation of residuals including the empirical Bayes estimates for the random effects. Altogether, we did not find any anomalies of note upon inspection of residuals. As an example, we provided plots of repeated measures residuals by both time and predicted value in Figure 4.3. In the ordered residual plot, we note bands at 0 and 90 days, common BHC assessment times, and otherwise note no discernable patterns. Although we see a striated diamond shape in the residuals vs. fitted plot, we assert this is largely due to the floor and ceiling of our repeated measure, which only takes integer values between 0 and 27, and the common predicted values falling towards the middle of this range as evidenced by the provided marginal histogram.

4.4 Discussion

We were able to demonstrate an adaptation and extension of a three-part shared random effects model to better evaluate clinical outcomes in the presence of both informative observation times and dependent termination when data is obtained from EHR. The extension included a method of estimating independent and dependent termination when such information goes unrecorded in medical records using an inverse cumulative hazard function. Both the three-part shared random effects model and the inverse cumulative hazard method can

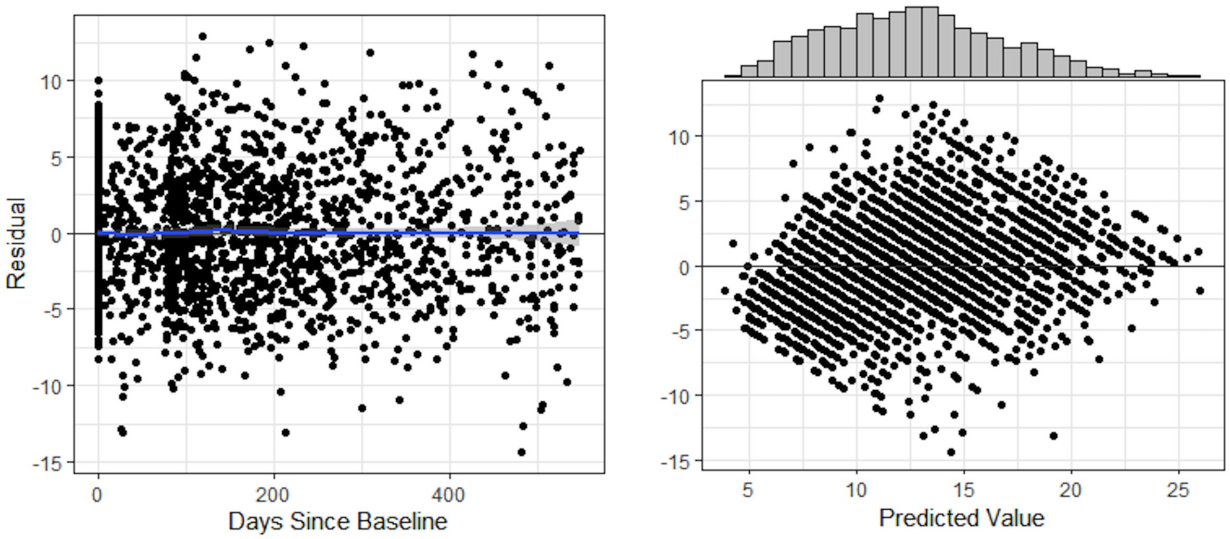


Figure 4.3: Ordered residual plot with LOESS fit to describe trend (left) and residual vs. fitted plot with corresponding marginal histogram of predicted values (right).

be implemented in SAS with basic assumptions that do not impact the ability to accurately interpret covariate effects on patients' outcome trajectories.

The selected application provided a straightforward example of a range of different customizations that can be incorporated into this model: approaches to account for unknown termination type and time, piecewise linear modeling of the repeated measures trajectory, piecewise constant baseline hazards, and incorporation of categorical and continuous covariates. The hope is for a wider-spread use of such models by applied researchers, with knowledge that they are flexible enough to meet the unique needs of any given research setting. As of 2017, 85.9% of office-based physicians in the United States used some form of electronic health or medical records [2]. As the adoption of EHR increases and their use becomes more systematic, it is expected that medical-based research will increasingly rely on data from these electronic sources. Given the susceptibility of EHR data to structural complexities, including issues of representativeness, considering and accounting for potential dependencies between patient observations and treatment termination when evaluating medical outcomes should be commonplace.

Analytical methods used to address these complexities should also focus on answering clinical questions, as primary interests regarding medical outcomes in EHR data often include the ability to draw inference on covariates. This shared random effects model, as well as use of the inverse cumulative hazard to predict independent and dependent termination, both allow for situation-specific flexibility in the specification of many model components including time scale, baseline hazard functions, functional forms of the trajectory, and longitudinal correlation structures, and the model output emphasizes evaluation of covariate effects. Moreover, the three-part shared random effects model has been shown to perform just as well as a reduced model (e.g. a model excluding informative observation times) when the reduced model is a valid fit to the data [41]. As such, we maintain that shared random effects models should be a standard analytical approach when considering EHR data as the tradeoff between ensuring unbiased parameter estimates and the cost to degrees of freedom typically favors the former in EHR data where sample size is often sufficiently large.

The presented application highlights the benefits of using this model to answer specific clinical questions and offers a direct extension to a recent short-term evaluation of response and remission from depression under BHA treatment [1]. Results offer a more holistic evaluation of the realized patient treatment experience in accounting for and understanding dependencies between visits to BHA, time under BHA treatment, and patients' depression symptoms. Specifying piecewise linear time allowed for consideration of symptom trajectories over clinically meaningful follow-up intervals that coincided with symptom reassessments. The form of the model also makes subsequent treatment-related questions, such as differences in dosing effects, treatment sessions, and medication management between patients with and without comorbid trauma, straightforward to evaluate.

4.5 Limitations and Future Extensions

We acknowledge several limitations in the current model implementation and recognize room for future work. In the application presented here, patients with only a baseline visit to BHA were excluded from the final analytical sample. We believe these patients are systematically different from those who did return for follow-up visits, but also believe that ultimately they are integral to an accurate description of the entirety of the BHA patient experience. The following chapter explores an extension of a joint model formulation, namely the inclusion of a cure fraction, to handle zero-inflation in medical records data. We also did not model the correlation structure for the longitudinal repeated measures. Future work using this application may benefit from considering this patient-level correlation.

Use of the inverse cumulative hazard to discriminate between patients who were dependently and independently terminated relies on estimating an inter-event time and using a cutoff to distinguish the two termination types. While we believe this method is effective, we are also interested in the comparative fit of a model that uses survival probabilities to model treatment termination in order to introduce additional flexibility and uncertainty in predicting dependent treatment termination. The following chapter, in addition to a cure fraction for zero-inflated data, explores the application of an adaptive cure frailty model to account for the probability of treatment termination after each successive visit rather than explicitly predicting treatment termination following the last-recorded patient observation.

Additionally, we know that using the two-part model in predicting independent and dependent termination has the potential for bias which could be addressed through an iterative approach. Future work might also include changes to the likelihood function to accommodate interval censoring which could be a valid alternative to assuming termination occurred one day after patients' last-recorded visit, although there are likely tradeoffs in terms of ease of computation.

Lastly, another extension involves the introduction of time-dependent covariates within

the longitudinal repeated measures process. The model notation already allows for such temporally dependent covariates, though implementation and interpretation would rely on careful consideration of within- and between-subject covariance specifications. While the use of baseline behavioral health conditions as fixed covariates does have clinical meaning, we expect that over the course of treatment, particularly extended periods of follow-up, changes in comorbid symptomatology will have a consequential impact on depression symptom trajectories. We will discuss time-dependent covariates in greater detail in Chapter 6.

Table 4.2: Parameter estimates from application of three-part shared random effects model to BHA data using the inverse cumulative hazard method to predict dependent termination.

Parameter	**Estimate (SE)**[1]	**95% CI**[2]	**p-value**
	Informative Observation Times		
$>$ 50% visits to a physician	-0.904 (0.031)	(-0.965, -0.843)	$<$0.001
	Repeated Measures		
intercept	11.777 (0.495)	(10.805, 12.749)	$<$0.001
Demographic Covariates			
gender	-0.499 (0.271)	(-1.029, 0.032)	0.066
age	0.011 (0.009)	(-0.007, 0.028)	0.230
marital status	-0.450 (0.276)	(-0.990, 0.091))	0.103
race/ethnicity	0.338 (0.260)	(-0.172, 0.847))	0.193
Behavioral Health Covariates			
PCL $>$ 50	2.935 (0.375)	(2.200, 3.670)	$<$0.001
GAD-7 $\geq$ 10	2.160 (0.302)	(1.567, 2.754)	$<$0.001
suicidal ideation	2.440 (0.266)	(1.917, 2.962)	$<$0.001
$>$ 50% visits to a physician	0.564 (0.260)	(0.054, 1.074)	0.030
Temporal Covariates			
time (days)	-0.051 (0.004)	(-0.058, -0.043)	$<$0.001
PCL $>$ 50 $\times$ time	-0.025 (0.007)	(-0.039, -0.011)	0.001
(time - 90)	0.042 (0.009)	(0.025, 0.059)	$<$0.001
PCL $>$ 50 $\times$ (time - 90)	0.054 (0.016)	(0.022, 0.087)	0.001
(time - 180)	0.011 (0.008)	(-0.004, 0.026)	0.143
PCL $>$ 50 $\times$ (time - 180)	-0.041 (0.015)	(-0.070, -0.012)	0.006
(time - 365)	-0.008 (0.007)	(-0.022, 0.006)	0.264
PCL $>$ 50 $\times$ (time - 365)	0.009 (0.013)	(-0.017, 0.035)	0.482
	Termination		
$>$ 50% visits to a physician	-0.569 (0.080)	(-0.726, -0.412)	$<$0.001
	Model Associations		
γ_1	2.184 (0.604)	(0.999, 3.368)	$<$0.001
γ_2	0.591 (0.173)	(0.251, 0.931)	$<$0.001
γ_3	-0.058 (0.022)	(-0.101, -0.016)	0.007
	Variance/Covariance		
σ_u^2	0.101 (0.009)	(0.082, 0.119)	$<$0.001
σ_v^2	7.473 (0.724)	(6.053, 8.894)	$<$0.001
σ_e	4.197 (0.075)	(4.049, 4.344)	$<$0.001

[1] SE: standard error
[2] CI: confidence interval

Table 4.3: Contrasts of symptom trajectories between changepoints from application of three-part shared random effects model to BHA data.

	0-3 Months		**3-6 Months**		**6-12 Months**		**> 12 Months**	
Contrast	**Estimate (SE)**	**p-Value**[1]	**Estimate (SE)**	**p-Value**	**Estimate (SE)**	**p-Value**	**Estimate (SE)**	**p-Value**
Slope: PCL $\leq$ 50	-0.051 (0.004)	<0.001	-0.009 (0.005)	0.103	0.002 (0.003)	0.435	-0.005 (0.005)	0.252
Slope: PCL > 50	-0.076 (0.006)	<0.001	0.020 (0.009)	0.022	-0.009 (0.005)	0.072	-0.008 (0.008)	0.313
Change in Slope[2]: PCL $\leq$ 50	0.042 (0.009)	<0.001	-0.030 (0.008)	0.143	-0.008 (0.007)	0.264	—	—
Change in Slope: PCL > 50	0.096 (0.014)	<0.001	0.011 (0.013)	0.020	0.001 (0.011)	0.895	—	—
Difference in Slopes	-0.025 (0.010)	<0.001	0.029 0.010	0.005	-0.012 (0.006)	0.052	-0.002 (0.009)	0.798

[1] p-values from F-tests with numerator DF = 1 and denominator DF = 947.
[2] Change assessed at end of designated interval (at 3-, 6-, and 12-months)

CHAPTER 5

Cure Models for EHR Data

5.1 Background

Medical and health-related research often involves count variables as outcomes of interest. Moreover, it is common in these applications for the proportion of zero-counts in the data to be greater than would be expected, leading to heavily right-skewed count data, often referred to as zero-inflated data. Many traditional analytic approaches rely on assuming data is normally distributed, but excess zeros remain problematic even after standard normalizing corrections like log-transformations, and nonparametric solutions, which often rely on ranks applied to outcome values, have problems with disproportionate ties among the clustered zero values [46]. Additionally, in many cases understanding and identifying covariates associated with zero values, or otherwise distinguishing zero from positive values, might hold particular clinical significance [47].

Zero-inflation does not only apply to standard count data, like defects in manufacturing [48], or continuous-value data, like medical costs associated with treatment for heart failure [49]. It also applies to survival and failure time data, where recurrent events are of interest, but a proportion of the study sample never experiences the recurrent event. In medical applications, the absence of any recurrent events could suggest a "cured" status, and only those patients who remain "non-cured" after that first event are at risk for future events. An example includes patients who receive surgery to remove a tumor, where tumor recurrence is the recurrent event and long-term follow-up without a tumor recurrence is the "cure".When

a significant proportion of patients are cured after some initial treatment, and thus no longer at risk for any future events, the population is then composed of a mixture of both cured and non-cured patients representing different disease or event susceptibilities. This mixture makes many survival analysis techniques, like Cox proportional hazards models, inapplicable as they tend to rely on an assumption of equal susceptibility across all persons and that with infinite follow-up time everyone will eventually experience the event of interest [50].

More generally, zero recurrent events can decompose into structural zeros and random zeros. A structural zero is otherwise termed a "cured" or an "unsusceptible", which means that even with an infinite wait time the event of interest will never occur, while a random zero, or a "non-cured" or "susceptible", means that the event of interest will eventually occur given a long enough wait time [51]. However, depending on the maximum follow-up time allotted, events among the non-cured may not be observed, with some subjects being censored before it is possible to determine whether they are truly cured or non-cured.

Cure models can estimate the probability of "cure" among a sample of patients and can help identify covariates associated with the probability of experiencing no recurrent events, sometimes referred to as long-term survival. Liu et al. (2016) [51] and Kim (2021) [52] demonstrate use of a logistic model that can be used to decribe the probability of cure alongside a joint model relating recurrent and terminal events among the non-cured. In these cases, the cure can occur after the baseline event, but non-cured individuals cannot transition into the cured group at a later time as there is no probability of a cure after subsequent recurrent events. Alternatively, Rondeau et al. (2013) [50] proposed a cure model capable of considering dependencies between the cure probability and the distribution of event times among the non-cured. They described a cure frailty model that uses random effects, much like the three-part shared random effects model described in Chapters 3 and 4, to account for within-cluster correlation among the "cured" and "non-cured", and a shared random effect between the cure probability and the frailty for event observation times such that patients have some probability of "cure" following each recurrent event.Thus, the cure

probability definition changes slightly to represent the probability of developing no future recurrent events after each given event, and I will refer to this formulation as an "adaptive cure frailty model". However, neither of these previous methods accounted for dependencies with a repeated measure, and each relied on explicitly observed cure events, distinguishable from dependent terminal events when applicable.

5.2 Defining a Cure Frailty Model for EHR Data

We propose both an extension and an application of a cure model: the addition of a cure fraction to the existing three-part shared random effects model to account for zero-inflated recurrent events data[51], and use of an adaptable cure frailty model[50] to account for unobserved treatment termination, essentially replacing the Cox proportional hazards model previously used for dependent terminal events. While existing literature does use cure frailties in shared random effects models, focus centers on evaluating survival and covariates associated with disease relapse and long-term cure. While this use is still relevant in applications to EHR data, there is the an additional interest in evaluating patient outcomes over time. It is easy to envision many clinical scenarios in which a cure fraction might depend not only on the intensity of recurrent event times, but also on some longitudinal patient outcome. Conversely, it is possible to imagine research interests that center around evaluation of patient outcome trajectories accounting for a cure probability. Cure models have been used to jointly evaluate survival times and longitudinal measurements[53,54]. However, our interest extends beyond single survival times to multiple recurrent event times, a distinct extension of these previous works.

Objectives in applying a cure frailty model to EHR data are twofold. First, a cure frailty model is a candidate solution to the issue of unobserved terminal events. Our initial solution involved estimating survival times between patients' last-recorded and next-hypothetical observation time and using this result classify dependent and independent termination. This

prediction method was an efficient means of modifying EHR data to fit a three-part shared random effects model with a Cox proportional hazards model for time spent under treatment. Benefits include the prediction of independent and dependent termination based on the intensity of previously recorded clinical observations through use of a two-part shared random effects model between observation times and repeated measures, and the ease of integration of the inverse cumulative hazard calculations into the three-part shared random effects model code. Moreover, we found the 85% correct classification rate in the simulation study to be satisfactory. However, this method has some notable limitations, foremost being its reliance on a two-part shared random effects model to obtain parameter estimates needed to calculate the survival times, even though it is already established that reduced models can yield biased estimates. An adaptive cure frailty model could offer an improvement in not relying on estimates from this two-part model, in addition to being an iterative and probabilistic approach.

The second objective in using a cure frailty model is its potential to address the zero-inflatedness of EHR data. It is possible to use a cure fraction in the existing three-part shared random effects model to account for the over-abundance of patients with zero recurrent events, and the adaptive cure frailty offers the possibility of accounting for both zero-inflatedness and treatment termination using the same model formulation. We will offer model notation for both a cure fraction for zero-inflated data, applied to the shared random effects model presented in Chapter 3, as well as an adaptive cure frailty model that allows for the possibility of "cure" after each successive recurrent event observation. We currently present the notation for the zero-inflated cure model and the adaptive cure model separately because of ongoing simulation studies and examination of convergence and stability within each model individually, but our ultimate goal is to combine notation into a single model that is flexible enough to handle a non-adaptive and/or an adaptive cure probability.

5.2.1 Model Notation for Zero-Inflated Recurrent Events

Using notation from Section 3.2 [16] and borrowing additional notation [50, 51] we let T_{ij} denote the j^{th} informative observation time for patient i, measured from their first observation (e.g. total time scale), where $i = 1, ..., n$ and $j = 1, ..., n_i$. We let X_i correspond to the total follow-up time and C_i represent the censoring time for patient i, measured from a baseline time $T_{i0} = 0$. We define δ_{ij} as a binary indicator for the observation of recurrent events such that $\delta_{ij} = 1$ if T_{ij} is observed and $\delta_{ij} = 0$ otherwise. More specifically, $\delta_{ij} = \mathrm{I}(T_{ij} < C_i)$, where $I(\cdot)$ is an indicator function. We continue to let $\boldsymbol{w_i^R}$, $\boldsymbol{z_i}$, and $\boldsymbol{w_i^C}$ be vectors of covariates for observation times, repeated measures, and dependent termination, respectively, and let $\boldsymbol{w_i^P}$ be vectors of covariates for the "cure" probability. While we present them as fixed-time covariates, it is possible to incorporate time-dependent covariates into this joint model formulation.

We continue to denote repeated measures by Y_{ij}. Repeated measures are observed in the presence of recurrent events but are not observed at every recurrent event. We also continue to assume that the absence of a repeated measure at a given recurrent event is independent of the longitudinal process itself such that conditional on the presence of a recurrent event the probability of observing a repeated measure is independent of Y_{ij}. In both of the following model formulations, we define two random effects, u_i and v_i. Both are assumed to be normally distributed and independent of each other, with $u_i \sim^{iid} N(0, \sigma_u^2)$ and $v_i \sim^{iid} N(0, \sigma_v^2)$. We also assume $e_{ij} \sim^{iid} N(0, \sigma_e^2)$.

To denote the "cure" status of patients we define another indicator A_i such that $A_i = 1$ if a patient is still "susceptible" ("non-cured") and will experience future recurrent events and $A_i = 0$ if the patient is no longer susceptible ("cured") or will not experience any more events. Thus we let $p_i = \mathrm{P}(A_i = 0)$ be the probability of "cure" for patient i and 1-p_i the probability of remaining "non-cured".

Cure models are often used with what is called a mixture population, or a population

comprised of both susceptible and non-susceptible persons, and the marginal survival function takes the general form of a mixture model:

$$S_t = p + (1-p)S(t|A=1) \tag{5.1}$$

We continue to use the frailty:

$$r_i(t) = r_0(t)exp(\boldsymbol{w_i^R}\beta + u_i) \tag{5.2}$$

to model the intensity of observation times (i.e. the timing of the recurrent events). The baseline hazard function for observation times is given by $r_0(t)$, and the cumulative baseline hazard by R_0. When using a cure model to account for zero-inflated recurrent events, we still allow for dependent terminal events, such as death or treatment termination, so we continue to use the proportional hazards model:

$$\lambda_i(t) = \lambda_0(t)exp(\boldsymbol{w_i^C}\eta + \gamma_2 u_i + \gamma_3 v_i) \tag{5.3}$$

to model the risk of dependent termination, where λ_0 represents the baseline hazard function, and Λ_0 the cumulative baseline hazard function, for depedent terminal events. Similarly, the notation for repeated measures remains consistent with that presented in Chapter 3:

$$y_{ij} = \boldsymbol{z_i}\alpha + \boldsymbol{t_{ij}}\kappa + \gamma_1 u_i + v_i + e_{ij} \tag{5.4}$$

Using the form of the general survival function above, we denote an overall survival function by:

$$S(t_{ij}|u_i) = p_i + (1-p_i) \times exp(-R_0(t_{ij}|A_i=1)exp(\boldsymbol{w_i^R}\beta + u_i)) \tag{5.5}$$

where the "cure" probability, p_i, is defined as:

$$p_i = P(A_i = 0) = \frac{exp(\boldsymbol{w_i^P}\phi)}{1 + exp(\boldsymbol{w_i^P}\phi)} \tag{5.6}$$

We have $\{\alpha, \beta, \eta, \phi\}$ as unknown parameters and the coefficients associated with the covariate vectors $\boldsymbol{z_i}$, $\boldsymbol{w_i^R}$, $\boldsymbol{w_i^C}$, and $\boldsymbol{w_i^P}$, respectively, and $\{\kappa\}$ is an unknown parameter and represents the coefficients associated with the longitudinal time vector $\boldsymbol{t_{ij}}$, the contents of which depend on the temporal specification. We continue to have unknown parameters $\{\gamma_1, \gamma_2, \gamma_3\}$ as the coefficients on the shared random effects between informative observation times and repeated measures, informative observation times and termination, and repeated measures and termination, respectively.

The cure probability $p_i(\cdot)$ represents the probability that patient i will never experience a recurrent event (after some baseline event), and only survival times until the first recurrent event are used in estimating the cure fraction. In a zero-inflated model, it is also important to note two assumptions that apply to "cured" individuals: they cannot experience any recurrent events and they cannot experience a dependent terminal event. This means that there remain four possible scenarios[51] the model must account for, as depicted by Figure 5.1. Note that the scenario in blue, which includes patients with no recurrent events and no dependent terminal event, is the only that includes "cured" patients as well as "non-cured" patients. The other three, denoted in green, include only "non-cured" patients.

We represent the relative likelihood contribution of each scenario as l_i^{1A}, l_i^{2A}, l_i^{3A}, l_i^{4A}, respectively, and the likelihood contribution of repeated measures as l_i^B, such that the conditional likelihood is:

$$\begin{aligned} L_i = (l_i^{1A})^{I(n_i>1,\Delta_i=1)} \times (l_i^{2A})^{I(n_i>1,\Delta_i=0)} \times \\ (l_i^{3A})^{I(n_i=1,\Delta_i=1)} \times (l_i^{4A})^{I(n_i=1,\Delta_i=0)} \times l_i^B \end{aligned} \tag{5.7}$$

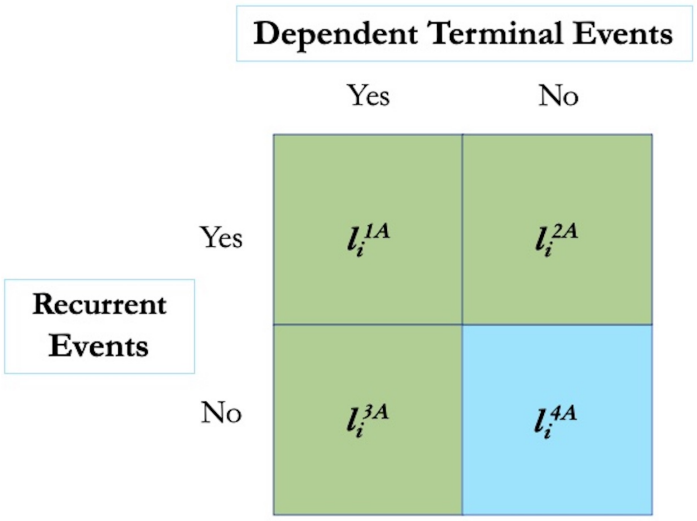


Figure 5.1: The four possible recurrent event and dependent terminal event scenarios to be accounted for in a zero-inflated cure model. The labels, l_i^{1A}, l_i^{2A}, l_i^{3A}, and l_i^{4A} correspond to the likelihood contributions in Equation 5.7.

where $I(\cdot)$ is an indicator function and $\Delta_i = I(D_i \leq C_i)$, and the marginal likelihood is:

$$L = \prod_{i=1} \int L_i \, f(u_i) \, f(v_i) \, du_i \, dv_i \tag{5.8}$$

where $f(u_i)$ and $f(v_i)$ are the density functions for the random effects u_i and v_i, respectively.

For ease of notation in defining the likelihood functions, I denote the survival function for recurrent events among the non-cured as:

$$S^R(t_{ij}|u_i, A_i = 1) = exp(-R_0(t_{ij})exp(\boldsymbol{w_i^R}\beta + u_i)) \tag{5.9}$$

Similary, we denote the survival function for dependent terminal events among the non-cured as:

$$S^D(x_i|u_i, v_i, A_i = 1) = exp(-\Lambda_0(t_{ij})exp(\boldsymbol{w_i^C}\eta + \gamma_2 u_i + \gamma_3 v_i)) \tag{5.10}$$

Using this notation, the likelihood functions for the four recurrent/terminal event scenarios described above are as follows:

1. Patient i with recurrent events and a dependent terminal event:

$$l_i^{1A} = (1-p_i) \times S_i^R(t_{ij}|u_i, A_i = 1) \times \prod_{j=1}^{n_i} r_i(t_{ij}|v_i, A_i = 1) \times \\ \lambda_i(x_i|v_i, A_i = 1)^{\Delta_i} \times S_i^D(x_i|u_i, v_i, A_i = 1) \tag{5.11}$$

   where $\Delta_i = I(D_i \leq C_i)$, an indicator for the presence of a dependent terminal event.

2. Patient i with recurrent events and no dependent terminal event:

$$l_i^{2A} = (1-p_i) \times S_i^R(t_{ij}|u_i, A_i = 1) \times \\ \prod_{j=1}^{n_i} r_i(t_{ij}|v_i, A_i = 1) \times S_i^D(x_i|u_i, v_i, A_i = 1) \tag{5.12}$$

3. Patient i with no recurrent events and a dependent terminal event:

$$l_i^{3A} = (1-p_i) \times S_i^R(t_{ij}|u_i, A_i = 1) \times \\ \lambda_i(x_i|v_i, A_i = 1)^{\Delta_i} \times S_i^D(x_i|u_i, v_i, A_i = 1) \tag{5.13}$$

4. Patient i with no recurrent events and no dependent terminal event:

$$l_i^{4A} = p_i + (1-p_i) \times S_i^R(t_{ij}|u_i, A_i = 1) \times S_i^D(x_i|u_i, v_i, A_i = 1) \tag{5.14}$$

The likelihood contribution for repeated measures for patient i is:

$$l_i^B = \frac{1}{(\sqrt{2\pi}\sigma_e)^{n_i}} \times exp[-\frac{1}{2\sigma_e^2} \sum_{j=1}^{n_i} e_{ij}^2] \tag{5.15}$$

where $e_{ij} = Y_{ij} - \boldsymbol{z_i}\alpha - \boldsymbol{t_{ij}}\kappa - \gamma_1 u_i - v_i$.

Note that while there are random effects shared between observation times, repeated measures, and terminal events, there is no random effect in the cure probability in this zero-inflated recurrent event formulation of the cure model.

5.2.2 Model Notation for Adaptive Cure Frailty

To model an adaptive cure frailty, or allow for the probability of a "cure" after each successive observation, we leave much of the notation the same as the above but ammend the definition of the cure status indicator such that $A_{ij} = 0$ if patient i is no longer susceptible ("cured") after the j^{th} informative observation time, and $A_{ij} = 1$ if patient i remains susceptible ("non-cured") after the j^{th} informative observation time. The cure probability notation also changes such that: $p_{ij} = \mathrm{P}(A_{ij} = 0)$ equals the probability of experiencing no more events after each successive event. Under this definition, the probability of a "cure" can change with time, and patients with recurrent events can still be a part of the "cured" subset. Note that we now include a random effect in the cure probability, and this random effect is shared with the recurrent event observation times frailty. We also continue to include a shared random effect between observation times and repeated measures, and could include another between repeated measures and the cure probability, but it is the shared effect between the cure probability and the observations process that allow for the adaptive cure model.

Using this notation, we now define the overall survival function, $S(t_{ij}|u_i, v_i)$, as:

$$S(t_{ij}|u_i, v_i) = p_{ij} + (1 - p_{ij}) \times exp(-R_0(t_{ij}|A_{ij} = 1)exp(\boldsymbol{w_i^R}\beta + u_i)) \tag{5.16}$$

The repeated measures process, denoted by y_{ij}, remains the same:

$$y_{ij}|(dN_{ij}(t) = 1) = \boldsymbol{z_i}\alpha + \boldsymbol{t_{ij}}\kappa + \gamma_1 u_i + v_i + e_{ij} \tag{5.17}$$

and the "cure" probability, denoted by p_{ij}, is now defined as:

$$p_{ij} = P(A_{ij} = 0) = \frac{exp(\boldsymbol{w_i^P}\phi + \gamma_2 u_i)}{1 + exp(\boldsymbol{w_i^P}\phi + \gamma_2 u_i)} \tag{5.18}$$

where $\{\alpha, \beta, \phi\}$ are unknown parameters and the coefficients associated with the covariate vectors $\boldsymbol{z_i}$,$\boldsymbol{w_i^R}$, and $\boldsymbol{w_i^P}$, respectively, and $\{\kappa\}$ is an unknown parameter and represents the coefficients associated with the longitudinal time vector $\boldsymbol{t_{ij}}$, the contents of which depend on the temporal specification.

In the above cure frailty model, γ_1 is the coefficient on the shared random effect between observation times and repeated measures and γ_2 the coefficient on the shared random effect between observation times and the "cure" probability.

The marginal likelihood function for the i^{th} patient is:

$$L_i = \int\int l_i^A \, l_i^B \, f(u_i) \, f(v_i) \, du_i \, dv_i \tag{5.19}$$

where $f(u_i)$ and $f(v_i)$ are the density functions for the random effects u_i and v_i, respectively.

The likelihood contribution for clinical observations for patient i is

$$\begin{aligned} l_i^A = \int_0^{x_i} \prod_{j=1}^{n_i} & [(1 - p_{ij})[r_0(t_{ij}|A_{ij} = 1)exp(\boldsymbol{w_i^R}\beta + u_i)] \times \\ & [exp(-R_0(t_{ij}|A_{ij} = 1)(exp(\boldsymbol{w_i^R}\beta + u_i))]]^{\delta_{ij}} \times \\ & \prod_{j=1}^{n_i} [p_{ij} + (1 - p_{ij})[exp(-R_0(t_{ij}|A_{ij} = 1)(exp(\boldsymbol{w_i^R}\beta + u_i))]]^{(1-\delta_{ij})} dt \end{aligned} \tag{5.20}$$

where δ_{ij} is an indicator of a clinical observation at time t_{ij}, and x_i is the total observed follow-up time for patient i. This likelihood function looks similar to the likelihood for observation times presented in Chapter 3, with the addition of the cure fraction notation.

The observation times now rely on some probability of remaining non-cured and continuing to be susceptible to future recurrent events.

The likelihood contribution for the repeated measures at associated observation times for patient i continues to be

$$l_i^B = \frac{1}{(\sqrt{2\pi}\sigma_e)^{n_i}} \times exp[-\frac{1}{2\sigma_e^2}\sum_{j=1}^{n_i} e_{ij}^2] \tag{5.21}$$

where $e_{ij} = Y_{ij} - \boldsymbol{z_i}\alpha - \boldsymbol{t_{ij}}\kappa - \gamma_1 u_i - v_i$.

During estimation, it may be necessary to adopt a zero-tail constraint [50]. The zero-tail constraint assumes that the survival function is "null" after the final observation, meaning that $S_0(t|A = 1) = 0$. This can improve the stability of maximum likelihood estimates when considering a cure fraction.

In this application, this "cure" probability represents all dependent termination. With this use, it is no longer necessary to explicitly generate termination "type" and "time" because the Cox proportional hazards model is effectively removed from the shared random effects model. However, in situations where it is possible to distinguish treatment termination for reasons related to symptom worsening from treatment termination for reasons related to symptom remission, it would be possible to use this adaptive cure frailty model and the Cox proportional hazards model simultaneously.

5.3 Simulation Study with Repeated Measures in the Presence of Zero-Inflated Recurrent Events, Informative Observation Times and Dependent Termination

The primary focus continues to be evaluating longitudinal clinical outcomes, so my first simulation study using a cure model is an extension of the zero-inflated model proposed

by Liu et al. (2016) [51]. We want to account for an overabundance of patients with a baseline observation only, while continuing to relate informative observation times, repeated measures, and dependent terminal events. We ran this simulation because existing literature using a cure fraction tends not to include repeated measures, so we wanted to confirm performance with the introduction of repeated measures and an additional shared random effect.

Referencing the model notation in Section 5.2.1 and borrowing some parameter specifications [51], we define z_i, w_i^R, w_i^C, and w_i^P to be the same fixed binary covariate at the individual level that can take values of 0 or 1 each with a probability of 0.5. We set β=1, η=1, $\phi = (-0.5, 1)^T$, and $\alpha = (1, 0.5)^T$, while we set $\kappa = 0.5$. We set the error term $e_{ij} \sim N(0, \sigma_e^2)$ with $\sigma_e^2 = 1$, and random effects $u_i \sim^{iid} N(0, \sigma_u^2)$ and $v_i \sim^{iid} N(0, \sigma_v^2)$ with $\sigma_u^2 = 1$ and $\sigma_v^2 = 1.0$. We assumed $(\gamma_1, \gamma_2, \gamma_3)^T = (1.0, 1.0, 1.0)^T$.

We used 2 + Uniform(0,6) to simulate an independent termination time (C_i) for each individual, and we simulated the informative observation times and dependent termination time (D_i) under a Weibull distribution. We defined the overall follow-up time for patient i as $X_i = \min(D_i, C_i)$, and the termination status as $event_i = 2$ if $X_i = D_i$ and $event_i = 1$ if $X_i = C_i$. For the baseline hazard function for observation times, we set the shape parameter to 1.25 and the scale parameter to 0.25, and for the baseline hazard function for dependent termination we set the shape parameter to 1.25 and the scale parameter to 0.10. In fitting the model in Proc NLMIXED, we used 5 quantiles to define the nodes of the piecewise constant baseline hazard functions for both observation times and terminal events.

For each individual, there is a baseline observation at time t_{i0}, and a repeated measure y_{i0} at time t_{i0}. We then compared the value of individual i's cure probability to a Uniform(0,1) random variable. Those whose cure probability exceeded the sampled uniform random variable were considered cured after their baseline observation and were deemed censored at C_i with $event_i = 1$. For those individuals whose cure probability was less than the sampled uniform random variable, we went on to generate a series of observation times until $t_{ij} >$

X_i, at which point we set termination time to X_i and termination status to $event_i$.

We simulated G=200 dataset replicates and results from fitting a zero-inflated three-part shared random effects model are found in Table 5.1. We note that including repeated measures with an additional shared random effect in a joint frailty model with a cure fraction for zero-inflated recurrent events data and a dependent terminal event maintains acceptible bias and coverage probabilities. Across the dataset replicates, the biases of the average recovered parameter estimates are 5.20% or less, while the coverage probabilities are all higher than 90%. We also acknowledge that we chose to use five quantiles to define the knots for the piecewise constant baseline hazard functions for informative observation and dependent termination times. We may have been able to decrease the biases and increase the coverage probabilities by a marginal amount if we used ten quantiles, but we determined that five was sufficient for this demonstration.

As a contrast, we attempted to fit a three-part shared random effects model ignoring the cure fraction to the same dataset replicates simulated under zero-inflated recurrent events conditions. The fit was so poor that Proc NLMIXED had difficulties converging.

Table 5.1: Simulation results from implementation of a three-part shared random effects model with use of a cure fraction to account for zero-inflated recurrent events data.

	Zero-Inflated Cure Model		
Parameter	**Estimate**[1]	**% Bias**[2]	**CP**[3]
	Informative Observation Times		
$\beta = 1$	0.998	0.20	95.50
	Repeated Measures		
$\alpha_0 = 1$	0.969	3.10	97.00
$\alpha_1 = 0.5$	0.487	2.60	96.00
$\kappa = 0.5$	0.496	2.60	95.50
$\sigma_e = 1$	0.998	0.20	94.00
	Termination		
$\eta = 1$	0.978	3.10	95.50
	Cure Probability		
ϕ_0 = -0.5	-0.508	1.60	95.50
$\phi_1 = 1$	1.008	0.80	96.00
	Model Association		
$\gamma_1 = 1$	0.995	0.50	94.50
$\gamma_2 = 1$	0.948	5.20	93.00
$\gamma_3 = 1$	0.985	1.50	95.00
	Variance/Covariance		
$\sigma_u^2 = 1$	0.971	2.90	94.00
$\sigma_v^2 = 1$	0.953	4.70	91.50

[1] Average parameter estimates across G=200 dataset replicates.

[2] Calculated as: abs((true parameter value - estimated parameter value) / true parameter value) x 100

[3] CP: coverage probability; calculated as: the percentage of 95% confidence intervals that contained the true parameter value.

CHAPTER 6

Application of Cure Model to BHA Data and Other Future Work

The work outlined in this dissertation builds upon previous uses of shared random effects models and offers extensions for use with EHR data, including approaches to handling specific complications inherent in data obtained from electronic records. It also provide a framework for both ongoing and future research applications.

6.1 Application of the Cure Model

6.1.1 Zero-Inflated Cure Model

In Chapter 4, we fit a three-part shared random effects model to EHR data obtained from BHA and found γ_1, γ_2, and γ_3, the coefficients on the shared random effects between the observation times, repeated measures, and terminal events processes respectively, to be statistically significant. This supports the hypothesis of both informative observation times and dependent termination in evaluating depression symptoms across follow-up among adult patients undergoing treatment in BHA.

However, one notable limitation in the construction of the data used for this application was the exclusion of patients who had only a baseline observation and no follow-up observations. We made the decision to exclude these patients because their inclusion in the analytical dataset caused model convergence problems and because it is ultimately believed patients

with only a baseline observation fundamentally differ from those patients with follow-up observations. Figure 6.1 illustrates the distribution of the total number of observations across adult patients ($\geq$ 18 years of age at the time of their first visit) under treatment in BHA who have a baseline visit after March 2013. We make the same restrictions to this sample as detailed in Chapter 4, with the distinct exception of excluding patients with only a baseline visit.

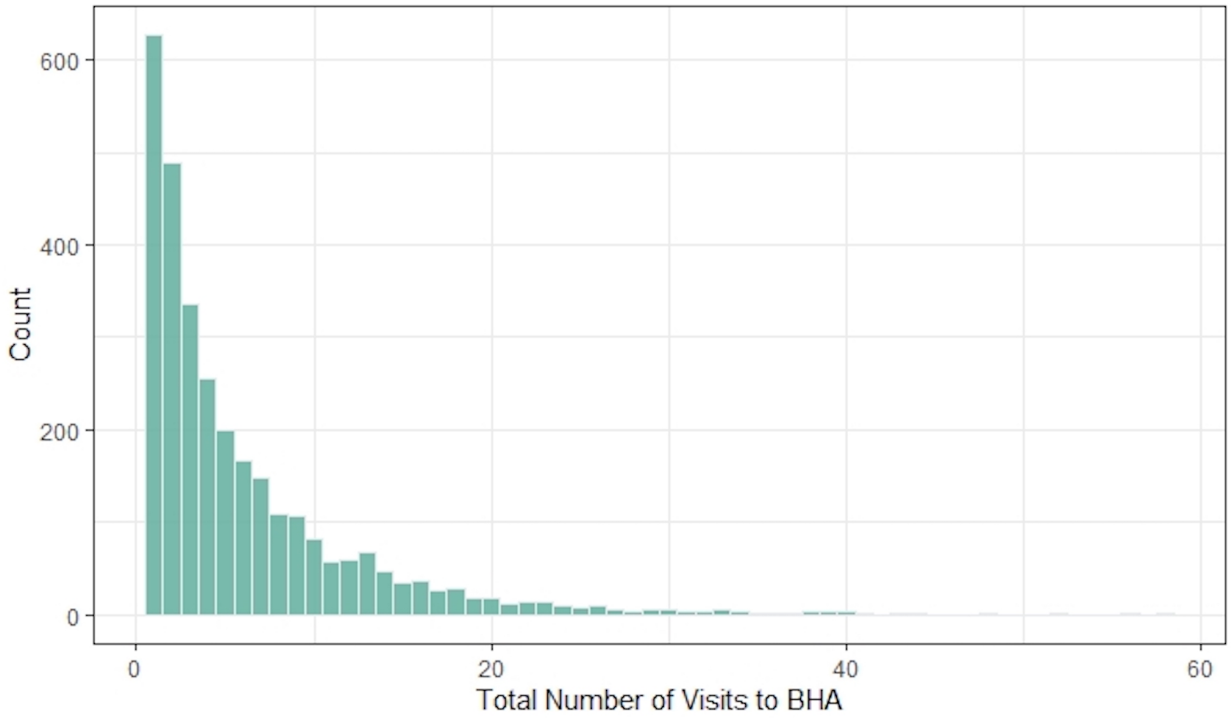


Figure 6.1: Total number of visits to a BHA clinic over the first 548 days of follow-up among patients with elevated PHQ-9 scores at baseline.

Note that 20.6% of patients have only one observation, which correlates with only a baseline visit and thus no recurrent events, as recurrent events are defined as follow-up visits to BHA. The total visit count distribution provides further evidence of the zero-inflatedness of the application data.

We also examined patient characteristics that are associated with having only a baseline observation in the sample. Among patients with an elevated PHQ-9 score at baseline,

preliminary two-sample tests between patients with and without follow-up visits to BHA suggest that provider type at baseline visit (physician vs. therapist; $p < 0.0001$), elevated trauma symptoms (PCL > 50; $p = 0.0079$), and the number of comorbid behavioral health coniditions at baseline ($p = 0.0189$) were associated with having only a baseline visit.

While we successfully completed a simulation study for a cure model applied to zero-inflated recurrent event data (described in Chapter 5), application of this model to EHR data from BHA has been met with some challenges. We believe the most substantial barrier surrounds the unobserved terminal events. One limitation of the inverse cumulative hazard method (described in Chapters 3 and 4) in estimating independent and dependent termination in the application data from BHA is its lack of variation in the classification and timing of termination for patients with only a baseline visit. Under the definition of dependent termination used for the application to BHA data, and using the inverse cumulative hazard method, the vast majority ($> 98\%$) of patients with only a baseline visit were labeled as dependently terminated with a termination time of 1 day (i.e. 1 day after their last-recorded observation).

In fitting a zero-inflated cure model, this result is challenging for a few reasons. The most prominent is the assumptions made in defining and constructing the likelihood functions, namely that those who were "cured" could not experience the dependent terminal event. Assuming that the N=627 patients with only a baseline visit represent a mixture of structural and random zeros, and thus a mixture of "cured" and "non-cured", a classification of dependently terminated is acceptible for the non-cured but violates the model assumptions among the cured. This leads to a discrepancy between the realized data and how the likelihood functions have been constructed to account for the overabundance of zero recurrent events.

We are actively working on an approach to handling unobserved treatment termination among those patients with only a baseline visit, particularly a method of differentiating between independent and dependent termination among this subset. We may ultimately arrive

at a solution that requires a sensitivity analysis to better understand how general we can make assumptions about treatment termination among the zero recurrent event group and still achieve model convergence, albeit with possibly conservative parameter estimates. In this particular application we have also considered expanding the analytical dataset further to include those patients without elevated PHQ-9 scores at baseline, though this change would require careful thought about the clinical implications of the results. The benefit of unrestricting the baseline sample would be the ability to better utilize number or severity of behavioral health conditions at baseline as a reference for distinguishing actual cure (e.g. no elevated symptoms on any behaivoral health measure) at baseline from dependent and independent treatment termination, unlike the application of the adaptive cure frailty model where the "cure" probability represents the probability of all dependent treatment termination.

6.1.2 Adaptive Cure Frailty Model

We believe that an application of an adaptive cure frailty model [50] could simultaneously account for zero-inflated recurrent event data while introducing a probabilistic approach to modeling treatment termination. In this setting, "cure" would refer to patients who had no more clinical visits and would effectively take the place of the proportional hazards model for treatment termination in the original analysis. Like the previous three-part model, it is not possible to distinguish between treatment termination for "positive" (e.g. treatment response, symptom remission) or "negative" (e.g. treatment non-compliance, referral to acute care) reasons.

We are continuing to work on a simulation study of the adaptive cure frailty model in the presence of informative observation times and repeated measures, using the cure probability to effectively model the probability of treatment termination after each successive visit. While we have confidence in the notation provided in Chapter 5, we have encountered some model convergence problems and parameter estimate instability when we attempt to apply

the model to the simulated data in Proc NLMIXED. Notably, we are able to recover the true parameter values for the covariate effects in the observation times and repeated measures processes, as well as the adaptive probability, with relative accuracy and acceptably high coverage probabilities. Overall, the coverage probabilities for all parameters range from 76%-92%, but we believe we can improve upon this estimation. The instability seems to be concentrated around the coefficients and variances of the shared random effects. We are working to determine the source of the instability, whether it be in the data simulation or in the model implementation, with our next step the application of the adaptive cure frailty model to the EHR data from BHA.

6.2 Additional Future Work

6.2.1 Shared Random Effects Model Diagnostics

We recognize the need for diagnostics methods for shared random effects models, and we intend to address this need in future work. Current literature, including work presented in this dissertation, makes use of existing diagnostics techniques used for hazards models or mixed effects linear models. While not inherently wrong, these methods were not developed for, and thus do not necessarily address, questions specific to shared random effects models.

Typically, diagnostic tools use some form of residuals calculated from fitted models, and much of the work on diagnostic tools for event and frailty models thus far involves martingale processes and residuals. Martingale residuals are often used as diagnostics tools for counting processes, relevant to the informative observation times process used in the three-part shared random effects models. Martingale residuals can be interpreted generally as "excess" events, or the difference between the observed and the expected number of events, given the model, over some time $[0,\ t]$ [55]. While there may exist more appropriate methods for any single diagnostic objective (e.g. assessing leverage compared to estimating a functional form), readily calculable residuals may be the key to accessible diagnostics techniques.

Martingale residuals have been used in applied research as part of diagnostics procedures for frailty models. Martingale residuals originally formed the basis of a goodness-of-fit test for Cox proportional hazards models, with a score test for the random effect in the hazard function [56]. The null hypothesis is that the variance of the random effect u_i in:

$$r_i(t) = r_0(t)exp(\boldsymbol{w_i^R}\beta + u_i) \tag{6.1}$$

is equal to zero, compared to the alternative hypothesis that the variance is greater than 0. In fact, we do evaluate the significance of the variances of the random effects included in the three-part models. For example, in Table 4.2, note that σ_u^2 and σ_v^2 are both statistically significant with $p < 0.001$ and 95% confidence intervals of (0.082, 0.119) and (6.053, 8.894), respectively.

Assuming $N_i(t)$ represents a counting process for the total number of events observed for patient i, with a standard-form cumulative intensity denoted by $\Gamma_i(t)$, this means $N_i(t)$ $= \Gamma_i(t) + M_i(t)$, where $M_i(t)$ is the martingale process associated with the relevant events process. Under this specification, and using standardized martingale residuals defined as:

$$M_i^*(t) = \frac{\hat{M}_i(t)}{\sqrt{\hat{var}\{N_i(t)\}}} \tag{6.2}$$

a plot of standardized martingale residuals over time t should be a flat line around a value of one as evidence of correct model specification [57]. While this method is a helpful visual tool, it is arguably best used as an informal graphical check due to strong associations between these estimated martingale residuals. Moreover, it only addresses model specification as far as marginally representing recurrent event observation times through use of an intensity process. Another graphical diagnostic technique relies on martingale residuals having uncorrelated increments, such that:

$$cov[M_i(t_0), M_i(t)] = var\{M_i(t_0)\} \tag{6.3}$$

for some $0 \leq t_0 < t$. It is possible to make use of this equality for testing specific model fit for continuous longitudinal data [58]. The left-hand side of the equality above, $cov[M_i(t_0), M_i(t)]$, can be evaluated at each measurement time and a plot of $cov[M_i(t_0), M_i(t)]$ against t should be a straight line with a slope of zero. Diggle et al. [58] also proposed a test statistic to accompany this covariance plot, but their specification of their response process and assumptions on their random effects differ slightly from the three-part shared random effects mode used in this dissertation.

As an extension of these graphical diagnostics methods, researchers focusing on recurrent event and longitudinal data proposed using empirical standard deviations of standardized martingale residual processes (SMRP) in plots against time, supplemented by plots of estimated martingale covariances:

$$\hat{C}(t) = n^{-1} \sum_i (\hat{M}_i(t_0)\hat{M}_t(t)) \tag{6.4}$$

with fit evidenced by no discernible trend over time [59].The authors note that the covariance procedure is robust to some of the issues frequently noted with using the SMRP procedure, but it still may not be possible to determine the type of misspecification and rather that these two graphical tools be used as something of an omnibus tool.

In addition to current diagnostics techniques only addressing parts of shared random effects models, many of them are not readily available in statistical software packages or rely heavily on user-defined code [60], as well as simulations or bootstrapping procedures. This, as with semi- or non-parametric formulations of shared random effects models, is a barrier to widespread use in standard practice, particularly among applied researchers. Our objective would be to continue to expand on this work by integrating diagnostics tools into a prepared code package requiring minimal input from the user (SAS would be preferrable for continuity with the model macro but we acknowledge R may be necessary), as well as developing methods to simultaneously assess each individual process and the overall fit of

the three-part shared random effects model.

6.2.2 Time-Dependent Covariates

Time-dependent covariates are generally defined as variables that change over a follow-up period and can be particularly applicable in medical applications where health-related measurements or comorbid medical conditions are repeatedly measured alongside survival, time-to-event, or a longitudinal outcome. Time-dependent covariates are often incorporated in survival and other time-to-event models, but less so in longitudinal conditional models, like the mixed effects model used for the repeated measures in the three-part shared random effects model. We are interested in exploring the inclusion of time-dependent covariates not only in the informative observation and dependent terminal events processes, but also the repeated measures process.

When time-dependent covariates are included in a mixed effects model, the model is typically adjusted to allow for both "within" and "between" effects resultant from these covariates [61]. This is to say that if a time-dependent covariate has variation within-subjects as well as between-subjects, the model formulation changes to allow for two distinct effects, which for a coefficient β and covariate x_{it} (e.g. covariate value for patient i at time t) decomposes as follows:

$$\beta_{x_{it}} \rightarrow \beta_W(x_{it} - \overline{x}_{i.}) + \beta_B \overline{x}_{i.} \tag{6.5}$$

where β_W represents the within-subject effect of the time-dependent covariate and β_B denotes the population-averaged effect of time-dependent covariate between subjects [62].

We presented shared random effects models using time-independent covariates, including comorbid baseline behavioral health conditions like elevated trauma symptoms and anxiety, as well as a series of demographic characteristics. While baseline behavioral health conditions do convey clinically important information, it is possible to imagine how behavioral

health treatment in BHA could simultaneously affect patients' depression and trauma symptoms, and how changes in those trauma symptoms could then have a differential impact on depression symptoms. There is also a variable that represents patients' provider type at each observation, with notably different treatments administered by therapists compared to physicians. We currently rely on a summary of provider information, majority provider (e.g. $> 50\%$ of visits to a physician), to represent and control for treatment type, but would consider a time-dependent indicator a more accurate covariate.

Difficulties in implementing time-dependent covariates in a longitudinal mixed effects model include identification of the exogeneity or endogeneity of the variables of interest. A time-dependent covariate is considered exogenous if it cannot be explained by other variables in the study (e.g. a "random" variable influenced by factors outside the system under study) and endogenous if it can be explained by other variables in the study, including the primary response variable in which the response and time-dependent covariate create somewhat of a feedback association [62]. It is also important to establish the level of exogeneity, generally divided in to four types, to ensure accurate model specification and interpretation [63,64], as well as consideration of within- and between-subject covariances.

6.2.3 Visit Pattern Identification

Ancillary to the proposed shared random effects models, we are interested in developing an event-time clustering algorithm to help identify patterns among patients' observation times. As evidenced in the application using data from BHA, there is significant inter- and intra-patient variation in clinical observation times, regardless of recommended treatment schedules. Additionally, there is some belief among clinicians that different observation patterns may differentially impact the effectiveness of behavioral health treatment. For example, more frequent clinical visits clustered towards the beginning of a patient's follow-up time with a gradual decrease in frequency after 3 months is thought to be more effective than the opposite or a patient with more sporadic visits but over a greater period of time. It

is also of interest to identify frequent and infrequent "users" of treatment, and what health-related variables affect this classification. There is also the possibility of using this kind of pattern identification, in addition to results obtained from the shared random effects model, in dynamic predictions. It would be of great clinical interest to be able to take a patient's baseline information, and information surrounding their first few visits, and predict what will happen to them over their stay in BHA treatment.

This is different from the way we have accounted for informative observation times through a shared random effects model. Results suggest that observation times are informative, associated with both the repeated outcome measure as well as the risk of dependent treatment termination. However, one distinct limitation of this model formulation is the assumption that within-subject random effects do not vary with time. For an example using the application to BHA, a patient's high concentration of visits to BHA in their first month of treatment would not affect their PHQ-9 score in the second month any differently than that same patient's sparse visits in their fifth month would affect their PHQ-9 scores in the sixth month. Essentially, this is because we have assumed that the association between repeated measures and observation times is constant across a patient's follow-up.

There are changes to the current three-part shared random effects model that could, at least partially, help mitigate this limitation, like further exploration of a random slope in the mixed linear effects model used for repeated measures, shared with the observation and/or termination time processes. While we did have success in simulation including a random slope effect shared between the repeated measures and termination processes, in the application work the addition of a random slope caused some model fit and convergence problems, perhaps due to overspecification of the model form given the realized data. We believe that pattern identification in visit times could supplement model results and provide tangible clinical feedback regarding treatment efficacy as reflected in patients' EHR.

CHAPTER 7

Conclusions, Implications, and Guidance

7.1 Summary and Conclusions

In this dissertation, I built upon previous work using shared random effects models and demonstrated methods required to apply such models to electronic health records data. Although EHR data is prone to the informative observation times and dependent terminal events these particular shared random effects models were developed to address, direct application of such models can be difficult, if not impossible, due to specific challenges not encountered in data from other sources, such as clinical trials or randomized controlled trials. I outlined methodologies and options for handling complications found in EHR data and provided tools for applied researchers to reasonably use these methods in clinical settings.

A particular structural complication, motivated by data obtained from UCLA Health's Behavioral Health Associates, was that of unobserved terminal events and unknown termination times. The available BHA data did not contain accessible information regarding referrals out to acute care, completion of treatment, or referrals back to primary care monitoring, nor notes pertaining to patients' perception of treatment. In considering patients' last-recorded observations in BHA there were no indications whether, if given infinite follow-up time, they would return for another visit or not, and if not then why not. While I was motivated by this specific application, it is a common occurrence in EHR for such informative details to go unnoted or remain contained in nonsystematic or inaccessible forms.

I considered the general idea of inverse transform sampling and proposed use of an inverse

cumulative hazard function, calculated from estimates obtained via fitting a two-part shared random effects model between informative observation times and repeated measures, to estimate survival times between patients' last-recorded and next-hypothetical observations. I made an assumption that if a patient's next-hypothetical visit time fell after the independent censoring time the patient was independently terminated, otherwise they were considered dependently terminated, but the method allows for situation-specific modifications of this assumption. This approach has multiple benefits, formost being patients' individual visit intensity and individual-level variation contributing to the time estimates, while accounting for dependencies with their outcome values. Moreover, the method fits seamlessly into the data structure and code needed to implement the primary three-part shared random effects model. Estimation of a specific survival time provides a hypothetical interval and allows for a sensitivity analysis based on choice of termination time within that interval.

I evaluated the performance of this proposed method through a simulation study in which I examined the bias and relative accuracy of parameter estimates recovered from fitting the three-part shared random effects model after using the proposed method to predict independent and dependent termination. I compared percentage biases and coverage probabilities to estimates obtained using known termination, those obtained via a naive prediction method, as well as those resultant from the often-used assumption that all termination is independent. I found that prediction of termination type using the proposed inverse cumulative hazard method produced estimates with minimal bias and high coverage probabilities when compared to results using known termination, with an average of 85% of termination types correctly classified based on simulated data designed to mimic the application. Compared to the naive prediction method, use of the inverse cumulative hazard favored misclassifying individuals as dependently terminatiod when they were truly independently terminated, which favorably produces more conservative results. Moreover, the proposed method resulted in notably more accurate estimates than treating all termination as independent, or ignoring termination altogether. After demonstrating this method's performance through

a simulation study, I applied it to EHR data from BHA and followed with an application of the three-part shared random effects model, where I further illustrated flexibility of the model formulation by considering a piecewise linear mixed effects model for patients' depression symptoms. Results included parameter estimates for evaluation of covariate effects on depression symptom trajectories over treatment follow-up as well as covariate effects on intensity of clinical observations and risk of dependent treatment termination. Estimates of coefficients on the shared random effects terms allowed for assessment of dependencies between clinical observation times, patients' PHQ-9 scores, and treatment termination, in which it was found that patients with a greater intensity of visits to BHA tended to have more severe depression symptoms, while patients with a greater intensity of visits to BHA and/or patients with less-severe depression symptoms had a greater risk of dependent treatment termination. Standard output from implementation in SAS Proc NLMIXED is in a recognizable form and incluses t-tests and p-values for significance tests on all parameter estimates, including the random effect variance parameters, the latter of which can help understand the fit and choice of the overall model. The application demonstrated the functionality of a customizable, practical, and effective solution to a common problem in naturalistic observational EHR data.

I then considered further extensions of shared random effects models through the examination of cure models. In addition to informative observation times and dependent termination, EHR is prone to zero-inflated recurrent events. In medical settings, there is often an overrepresentation of individuals with only a baseline event or clinical observation where a "cure" precludes future recurrent events. I expanded on previous use of cure models, which focused on joint time-to-event and survival data, to include repeated measures and emphasize understanding how different processes continue to affect an outcome measurement over time. I performed a simulation study with zero-inflated recurrent events data and fit a three-part shared random effects model with a cure fraction at baseline, and have set the framework for applying the same model to EHR data from BHA. I addressed additional

complexity in the use of a cure fraction at baseline with unobserved terminal events and offered possible solutions. I also explored the possibility of a cure model offering a probabalistic approach to unobserved treatment termination, and how an adaptive cure frailty model has the ability to simultaneously address zero-inflatedness and the probability of treatment termination. Preliminary simulation studies using an adaptive cure frailty model with informative observation times and repeated measures resulted in low percentage biases and acceptable coverage probabilities for some parameter estimates, but I believe I can improve upon the stability and convergence of the model. Ultimately, I intend to combine notation from the zero-inflated cure model and the adaptive cure frailty model such that a singular shared random effects model, with use of appropriate indicator functions, could account for "cure" after baseline and/or "cure" after each recurrent event, or neither (i.e. the original three-part model).

Another primary objective was to make the three-part shared random effects model easily implementable by applied researchers. The proposed methods focus on providing analytical solutions to realized data with the goal of addressing clinical questions, so it was important to develop methodologies with the smallest-possible barriers to practical use in clinical settings. Publicly available code and instructions on implementation of shared random effects models were either not available or inaccessible to an applied researcher. Even within the literature that used SAS, there lacked sufficient guidance on data setup requirements or explanation of the crucial inputs necessary to successfully run the model in Proc NLMIXED. The user-specified loglikelihod functions require nonintuitive data preparation, and Proc NLMIXED offers many options and possible inputs although only a subset are relevant. In developing the macro, I am able to provide a complete tool for researchers evaluating a longitudinal outcome in the presence of informative observation times and/or dependent termination. Importantly, the macro is also designed to accommodate data where existence of a dependent terminal event is unknown, a common feature in EHR data applications. By detailing which inputs the user must supply and including the detailed data preparation within the macro, I have

provided a tangible tool for applied researchers.

7.2 Using EHR Data for Research

Many of the challenges addressed in this dissertation, both methodologically and in preparing the application dataset for use, emphasize some of the ways EHR data is not optimized for use in research. The inherent value of EHR data is substantial, and while there have been improvements in data structure and quality since their inception, it remains that electronic health records were developed to expedite billing and insurance claims and support clinical care. Research is a secondary, albeit increasingly consequential, use.

For example, both the inverse cumulative hazard method for estimating dependent and independent treatment termination, and the application of an adaptive cure frailty model for a probabilistic approach to dependent treatment termination, are proposed solutions to rectify the lack, or inaccessibility, of particular information from EHR. With the increasing reliance on electronic records in medical systems, and the promise of harnessing EHR to provide more effective patient care, these methods are necessary to efficiently using EHR data.

Streamlined methods of linking different electronic databases, whether within the same health system and/or between health systems, would make it possible to ascertain engagement in healthcare before and after a given health-related event of interest. Additionally, including systematic fields in EHR that queried patient referral information would improve the determination of end-of-care for a given medical episode. Taken together, researchers would be able to draw inferences between treatment termination and subsequent engagement in medical care. Similarly, systematic fields establishing treatment recommendations would make it possible for researchers to determine deviations from intended care. "Systematic fields" is an important component of many recommendations for improvement. It is possible that some of the abovementioned information is contained somewhere in EHR, but if it is

generally inaccessible, like embedded in clinical notes, its use becomes impractical in most applied research environments.

Ideally, researchers and clinicians alike want to be able to evaluate data from EHR and assess treatment effectiveness. Certain structural complications, such as informative observation times and dependent terminal events, are unavoidable, which is why practical analytical tools that make it simpler to account for such factors are crucial to effective use of EHR data in clinical and applied research settings. Small, efficient changes to the way EHR data is collected, such as a wider breadth of systematic fields, and referral and treatment recommendations, could circumvent any additional methods developed only to correct for the lack of information.

7.3 Implementing a Shared Random Effects Model in SAS

7.3.1 Example Execution of SAS Macro

We wrote a SAS macro that not only implements the three-part shared random effects model but also the two-part model followed by inverse transform sampling for estimating dependent/independent termination described in Section 3.4. This was a crucial part of our effort to demonstrate the functionality of this model in an applied setting. We wanted to provide an analytical tool that required minimal input from the user to encourage wider use, particularly in clinical environments featuring EHR data. The full macro code can be found in the Appendix.

The execution of the SHARED3_RANEFF macro below applies the methods detailed in this paper to simulated data contained within a SAS dataset named SIMDAT_1. Through specification of mandatory arguments below, SAS will fit a model where repeated measures of the continuous outcome are contained within Y, visit date is contained within DAY, PID is used to uniquely identify individuals, and Z is a dichotomous covariate included in all three model components. Y, DAY, PID, and Z are all variables present in the SIMDAT_1

SAS dataset. The user must also supply another variable from the referenced SAS dataset which contains a date for each individual. The date provided will be taken to be the assumed termination date if a dependent terminal event is determined for the individual based on the inverse cumulative hazards approach. For example, in the BHA application this variable would consist of the date one day after the patient's last visit date for each patient in the dataset. In SIMDAT_1, this variable is named TERM. The execution below assumes 10 pieces for the piecewise baseline hazard functions for both informative observation times and dependent terminal event times and enforces censoring at 360.64 days after each individual's baseline visit. Other arguments are used to specify starting values and a minimum number of iterations, both of which are fed to both the NLMIXED procedures in sequence (corresponding to the two- and three-part models). The example below does not assume any splines but the SHARED3_RANEFF does have the flexibility to specify a spline model.

Table 7.1: An excerpt from the SIMDAT_1 csv file, used in the example execution of the SAS macro SHARED3_RANEFF.

PID	Z	DAY	Y	TERM
1	0	0	6.631677	280.3032
1	0	181.266	46.07549	280.3032
2	0	0	10.73156	884.8686
2	0	20.76064	13.24052	884.8686
2	0	106.6951	29.92867	884.8686
2	0	244.3545	58.32861	884.8686
2	0	245.4267	58.85101	884.8686
2	0	253.2158	58.6007	884.8686
3	0	0	8.133787	178.9796
4	1	0	9.361513	126.3453
4	1	62.73256	21.10914	126.3453
4	1	89.0846	26.41639	126.3453

The excerpt from the SIMDAT_1 example dataset demonstrates the general format required to call the SAS macro (Table 7.1). Importantly, note that the data is in a "long" format; there is a separate row in the dataset for each time an individual is observed (each observation time), starting with a baseline observation.

Figure 7.1: Example execution of the SHARED3_RANEFF macro.

```
%SHARED3_RANEFF(DAT=SIMDAT_1, Y=Y, TIME=DAY,
        UNIQ_ID=PID, TIME_END=360.64, TIME_TERM=TERM,
        COV_LONG = Z,
        COV_RECR = Z,
        COV_TERM = Z,
        NUMB_RECR = 10,
        NUMB_TERM = 10,
        DURATION_END = 360.64,
        MINITER = 250,
        SV_RECR = 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01,
        SV_TERM = 0.001 0.001 0.002 0.003 0.003 0.004 0.006 0.004 0.008 0.0012,
        SV_INTERCEPT = 10,
        SV_TIME = 0.2,
        SV_COV_LONG = 1,
        SV_BETA = 1,
        SV_ETA = 1,
        SV_GAMMA = 1.5 -0.5 1,
        SV_VAR = 1 0.5 1);
```

Figure 7.1 displays the example call of the SAS macro. In addition to the variables detailed in the previous paragraphs, there are some other values that a user must provide. MINITER is the minimum number of iterations the NLMIXED procedure will attempt, and choice of this value depends on a few factors. In this example call, it is set at 250 iterations, which was sufficient for application to simulated data where realistic starting values could be easily specified due to the known parameter values. However, there were instances when using the application data from BHA that I required far more iterations, sometimes upwards of 1000, to ensure that the resulting parameter estimates were accurate, and the model convergence was stable. This number might vary based on the user's confidence in their starting values for their model parameters, the size of the dataset, and the model convergence information provided by Proc NLMIXED.

The remaining values, represented by the last nine lines of the SHARED3_RANEFF macro call, are starting values for model parameters. Starting values are ultimately very important to the fit and convergence of the NLMIXED procedure and can influence the model convergence time as well as whether Proc NLMIXED encounters any warnings or errors during its run.

7.3.2 Selecting Starting Values for Model Parameters

An important component of implementing this model in Proc NLMIXED is the selection and specification of starting values for all unknown parameters. This includes the piecewise constant baseline hazard functions for informative observation times (SV_RECR) and dependent terminal event times (SV_TERM), the intercept term (SV_INTERCEPT), time variables (SV_TIME), and other covariates (SV_COV_LONG) in the repeated measures model, the covariates for the informative observation times (SV_BETA) and terminal events (SV_ETA) processes, and finally the coefficients on the shared random effects terms (SV_GAMMA) as well as the variances of the shared random effects and individual error term (SV_VAR). If no starting value is specified, the SAS default is to assign a value of 1, which can be a very

poor choice for many parameters. In fact, depending on the convergence criteria used and the distance this default value is from the true value of the estimate, poorly defined starting values can cause the model to fail to converge.

The most straightforward approach to determining starting values is to run simple models that disregard the shared random effects. For example, this would mean using SAS Proc MIXED to run a longitudinal mixed effects model for repeated measures (ignoring informative observation times and dependent terminal events), including all relevant covariates, and using the resultant parameter estimates as starting values in SV_INTERCEPT, SV_TIME, and SV_COV_LONG. The same applies to implementing individual models for informative observation times and terminal events. At the very least, these parameter estimates can be combined with a higher number of iterations of Proc NLMIXED that may slow the initial runtime of the model but will yield values for most parameters that can be utilized in further model implementations. This is the general approach I followed in the application of the three-part shared random effects model to BHA data.

Other helpful information, including common warnings and errors produced by Proc NLMIXED, along with potential solutions, were detailed by Kiernan, Tao, & Gibbs (2012) [65]. Many of their proposed approaches were helpful for troubleshooting in the early stages of my adaptation of a shared random effects model for EHR data, particularly in understanding the importance of supplying starting values.

7.4 Benefits of Shared Random Effects Models

Given that informative observation times and dependent terminal events are inherent to EHR data, regardless of potential changes to EHR to make it more optimal for use in research, it is important to have practical and usable analytical solutions. There are multiple reasons why a three-part shared random effects model is an ideal candidate for EHR data, particularly in clinical settings. Though mentioned throughout this dissertation, the following bullet points

summarize the primary benefits of using a shared random effects model:

- **A parametric specification allows for implementation in SAS Proc NLMIXED**:
  - NLMIXED is an existing procedure within SAS with a relatively low programmatic barrier. It allows for user-specification of loglikelihood functions, which increases the distributional flexibility of models it can accommodate. The output is also in a recognizable form, and includes parameter estimates, standard errors, 95% confidence intervals, and p-values for statistical significance. A SAS macro, requiring minimal user inputs and implementing the described three-part shared random effects model using an inverse cumulative hazard function to estimate termination, is included in the Appendix.
- **The model formulation puts an emphasis on interpretation of covariates**:
  - Covariates can be included in the repeated measures, informative observation times, and dependent terminal event processes. In a linear mixed effects model for the repeated measures, interpretation follows the recognizable form: "A ... unit change in x is associated with a ... unit change in y", which makes application of the results relatively straightforward. Because the focus is on the interpretation of covariates, it is advantageous to treat the baseline hazard functions as nuisance parameters, because approximation of the true underlying distribution (e.g. using piecewise constant baseline hazard functions) is sufficient.
- **The mixed effects model for repeated measures allows for flexibility of the time specification**:
  - The model formulation can accommodate linear, quadratic, piecewise linear (as was demonstrated in this dissertation in the application to BHA data) temporal specifications. There is also the possibility to include correlation structures, such

as autoregressive or compound symmetry, that might be useful for longitudinal modeling.

- **With adequate sample size (as is usually the case in research involving EHR), assuming dependencies between observation times, outcomes, and terminal events is a conservative approach**:
  - An assumption of dependencies, even if there are none, will at worst result in non-significant estimates of coefficients on the shared random effects terms, and a loss of some degrees of freedom, but there is nothing inherently biasing about using a three-part shared random effects model when a reduced model would have been sufficient.

For these reasons, a shared random effects model is a practical approach for use in clinical settings where the primary interest is understanding outcome trajectories while considering covariate effects. Because of the tendency of EHR data to contain structural complexities, like informative observation times and dependent terminal events, use of a three-part shared random effects model as described in this dissertation should be more commonplace.

APPENDIX A

Macro to Implement Shared Random Effects Model

```
/** MANDATORY ARGUMENTS **/

*DAT = DATASET NAME;
*Y = COLUMN NAME IN DAT CORRESPONDING TO THE OUTCOME THAT WILL BE
    MODELED LONGITUDINALLY;
*TIME = COLUMN NAME IN DAT CORRESPONDING TO TIME;
*UNIQ_ID = COLUMN NAME IN DAT CORRESPONDING TO UNIQUE ID FOR EACH
    INDIVIDUAL;
*TIME_END = VALUE CORRESPONDING TO THE TIME ALL RECORDS WOULD HAVE
     BEEN CENSORED (DATA EXTRACTIION DATE/TIME);
*TIME_TERM = VALUE CORRESPONDING TO THE ASSUMED TERMINAL EVENT
    TIME FOR AN INDIVIDUAL IF A TERMINAL EVENT IS DETERMINED TO
    HAVE HAPPENED;

/** OPTIONAL ARGUMENTS **/

*KNTS_SPLINE = LIST CONTAINING THE LOCATION OF THE SPLINE KNOTS (
    IF EMPTY, ASSUMES NO KNOTS). LOCATION SHOULD BE IN TERMS OF
    LENGTH OF TIME SINCE FIRST RECORD FOR EACH INDIVIDUAL;
*COV_LONG = LIST OF COLUMN NAMES IN DAT CORRESPONDING TO THE
```

```
    COVARIATES THAT WILL BE INCLUDED AS INTERCEPT TERMS IN THE
    LONGITUDINAL COMPONENT OF THE MODEL (IF EMPTY, ASSUMES NO
    COVARIATES);
*COV_LONG_TIME = LIST OF COLUMN NAMES IN DAT CORRESPONDING TO THE
    COVARIATES THAT WILL BE INCLUDED AS SLOPE TERMS IN THE
    LONGITUDINAL COMPONENT OF THE MODEL (IF EMPTY, ASSUMES NO
    COVARIATES);
*COV_RECR = LIST OF COLUMN NAMES IN DAT CORRESPONDING TO THE
    COVARIATES THAT WILL BE INCLUDED IN THE RECURRENT EVENTS
    COMPONENT OF THE MODEL (IF EMPTY, ASSUMES NO COVARIATES);
*COV_TERM = LIST OF COLUMN NAMES IN DAT CORRESPONDING TO THE
    COVARIATES THAT WILL BE INCLUDED IN THE TERMINAL EVENTS
    COMPONENT OF THE MODEL (IF EMPTY, ASSUMES NO COVARIATES);
*NUMB_RECR = NUMBER OF PIECES FOR THE PIECEWISE BASELINE HAZARD IN
    THE RECURRENT EVENTS COMPONENT OF THE MODEL (DEFAULTS TO 5);
*NUMB_TERM = NUMBER OF PIECES FOR THE PIECEWISE BASELINE HAZARD IN
    THE TERMINAL EVENTS COMPONENT OF THE MODEL (DEFAULTS TO 5);
*DURATION_END = VALUE CORRESPONDING TO THE DURATION OF TIME AFTER
    WHICH CENSORING IS ASSUMED (DEFAULTS TO 365);
*MINITER = MINIMUM NUMBER OF ITERATIONS TO BE SPECIFIED FOR THE
    NLMIXED PROCEDURES (DEFAULTS TO 250);
*SEED = ENSURES REPLICABILITY (DEFAULTS TO 202401);

/** STARTING VALUE ARGUMENTS (MANDATORY DEPENDING ON MODEL
    SPECIFICATIONS)**/

*SV_RECR = LIST OF STARTING VALUES FOR THE PIECEWISE BASELINE
```

HAZARD IN THE RECURRENT EVENTS COMPONENT OF THE MODEL (LENGTH SHOULD EQUAL NUMB_RECR);

*SV_TERM = LIST OF STARTING VALUES FOR THE PIECEWISE BASELINE HAZARD IN THE TERMINAL EVENTS COMPONENT OF THE MODEL (LENGTH SHOULD EQUAL NUMB_TERM);

*SV_INTERCEPT = STARTING VALUE FOR THE INTERCEPT IN THE LONGITUDINAL COMPONENT OF THE MODEL (LENGTH SHOULD EQUAL 1);

*SV_TIME = LIST OF STARTING VALUES FOR THE LONGITUDINAL COMPONENT OF THE MODEL CORRESPONDING TO TIME (FIRST IN LIST) AND SPLINE KNOTS IN THE ORDER SPECIFIED IN KNTS_SPLINE (LENGTH SHOULD EQUAL LENGTH OF KNTS_SPLINE PLUS 1);

*SV_COV_LONG = LIST OF STARTING VALUES FOR THE LONGITUDINAL COMPONENT OF THE MODEL CORRESPONDING TO COVARIATES LISTED IN COV_LONG (LENGTH SHOULD EQUAL LENGTH OF COV_LONG);

*SV_COV_LONG_TIME = LIST OF STARTING VALUES FOR THE LONGITUDINAL COMPONENT OF THE MODEL CORRESPONDING TO TIME AND COVARIATES LISTED IN COV_LONG_TIME (LENGTH SHOULD EQUAL LENGTH OF COV_LONG_TIME TIMES (LENGTH OF KNTS_SPLINE PLUS 1));

*SV_BETA = LIST OF STARTING VALUES FOR THE RECURRENT EVENTS COMPONENT OF THE MODEL CORRESPONDING TO COVARIATES LISTED IN COV_RECR (LENGTH SHOULD EQUAL LENGTH OF COV_RECR);

*SV_ETA = LIST OF STARTING VALUES FOR THE TERMINAL EVENTS COMPONENT OF THE MODEL CORRESPONDING TO COVARIATES LISTED IN COV_TERM (LENGTH SHOULD EQUAL LENGTH OF COV_TERM);

*SV_GAMMA = LIST OF STARTING VALUES FOR GAMMA1, GAMMA2, AND GAMMA3 (IN THAT ORDER);

*SV_VAR = LIST OF STARTING VALUES FOR VARIANCE(U), VARIANCE(V),

```
    AND VARIANCE(E) (IN THAT ORDER);


%MACRO SHARED3_RANEFF(DAT, Y, TIME, UNIQ_ID, TIME_END, TIME_TERM,
    KNTS_SPLINE=, COV_LONG=, COV_LONG_TIME=, COV_RECR=, COV_TERM=,
    NUMB_RECR=5, NUMB_TERM=5, DURATION_END=365, MINITER=250, SEED
    =202401,
                              SV_RECR=, SV_TERM=, SV_INTERCEPT=, SV_TIME=,
                                  SV_COV_LONG=, SV_COV_LONG_TIME=, SV_BETA
                                 =, SV_ETA=, SV_GAMMA=, SV_VAR=);

%LET NUMB_KNTS = %SYSFUNC(COUNTW(%STR(&KNTS_SPLINE)));
%LET NUMB_COV_LONG = %SYSFUNC(COUNTW(%STR(&COV_LONG)));
%LET NUMB_COV_RECR = %SYSFUNC(COUNTW(%STR(&COV_RECR)));
%LET NUMB_COV_TERM = %SYSFUNC(COUNTW(%STR(&COV_TERM)));
%LET NUMB_COV_LONG_TIME = %SYSFUNC(COUNTW(%STR(&COV_LONG_TIME)));

PROC SORT DATA=&DAT; BY &UNIQ_ID. &TIME; RUN;
DATA DAT1;
 SET &DAT;
 RETAIN FIRST_TIME CTIME TTIME;
 BY &UNIQ_ID;
 IF FIRST.&UNIQ_ID THEN DO;
  FIRST_TIME = &TIME;
  CTIME = &TIME_END - &TIME;
  TTIME = &TIME_TERM - &TIME;
  T = 0;
```

```
 END;
 ELSE T = &TIME - FIRST_TIME;
 IF T <= &DURATION_END;
 DROP FIRST_TIME;
RUN;

PROC SORT DATA=DAT1; BY &UNIQ_ID DESCENDING &TIME; RUN;
DATA DAT1;
 SET DAT1;
 RETAIN LAST_TIME;
 BY &UNIQ_ID;
 LAST = 0;
 IF FIRST.&UNIQ_ID THEN DO;
  LAST_TIME = T;
  LAST = 1;
 END;
 DI = TTIME;
 CI = MIN(&DURATION_END,CTIME);
 IF T <= &DURATION_END;
 DROP LAST_TIME CTIME;
RUN;
PROC SORT DATA=DAT1; BY &UNIQ_ID &TIME; RUN;

%IF (&NUMB_KNTS NE 0) %THEN %DO;
DATA LONG;
 SET DAT1;
 ARRAY TIME_SPLINE_{&NUMB_KNTS};
```

```
 %DO I=1 %TO &NUMB_KNTS;
  VAL = %SCAN(&KNTS_SPLINE,&I);
  IF T <= VAL THEN TIME_SPLINE_{&I} = 0;
  ELSE IF T > VAL THEN TIME_SPLINE_{&I} = (T - VAL);
 %END;
 IF Y NE .;
 DROP VAL;
RUN;
%END;
%ELSE %DO;
DATA LONG;
 SET DAT1;
 IF Y NE .;
RUN;
%END;

DATA RECR;
 SET DAT1 (WHERE=(T > 0));
RUN;

%IF %SYSFUNC(MOD(100,&NUMB_RECR))= 0 %THEN %DO;
 %LET INCREMENT_RECR = %EVAL(100/&NUMB_RECR);
%END;
%ELSE %DO;
 %LET INCREMENT_RECR = %EVAL((100/&NUMB_RECR)+ 1) ;
%END;
```

```
%IF %SYSFUNC(MOD(100,&NUMB_TERM))= 0 %THEN %DO;
 %LET INCREMENT_TERM = %EVAL(100/&NUMB_TERM);
%END;
%ELSE %DO;
 %LET INCREMENT_TERM = %EVAL((100/&NUMB_TERM)+ 1) ;
%END;

PROC UNIVARIATE DATA=RECR NOPRINT;
 VAR T;
 OUTPUT PCTLPRE=R_ PCTLPTS=0 TO 99 BY &INCREMENT_RECR,100 OUT=
     PCTL_RECR;
RUN;

/*** FITTING THE TWO-PART MODEL ***/

DATA DAT2;
 IF _N_ = 1 THEN SET PCTL_RECR;
 SET DAT1;

 ARRAY QUANT_R {*} R_:;
 ARRAY DUR_R {&NUMB_RECR};
 ARRAY EVENT_R {&NUMB_RECR};

 DO I=1 TO &NUMB_RECR;
  DUR_R{I} = 0;
  EVENT_R{I} = 0;
 END;
```

```
  IF  LAST=0  THEN DO;
   DO  I=2  TO  (&NUMB_RECR  +  1);
     IF  T <= QUANT_R{I}  THEN DO;
      EVENT_R{I-1} = 1;
      I  =  (&NUMB_RECR  +  1);
     END;
   END;
  END;

  ELSE DO;
   DO  I=2  TO  (&NUMB_RECR  +  1);
     IF  T <= QUANT_R{I}  THEN DO;
      DUR_R{I-1} = MAX(T-QUANT_R{I-1},0);
           EVENT_R{I-1}=1;
      I  =  (&NUMB_RECR  +  1);
     END;
     ELSE DUR_R{I-1} = QUANT_R{I}-QUANT_R{I-1};
   END;
  END;
  DROP  I;
RUN;

DATA NLMIXED_TWO;
  SET LONG  (IN=IND1)
       DAT2  (IN=IND2);
  IF  IND1  THEN DO;
```

```
    OUTC = Y;
    OUTC_TYPE = 1;
  END;
  ELSE IF IND2 THEN DO;
   OUTC = T;
   OUTC_TYPE = 2;
  END;
RUN;

PROC SORT DATA=NLMIXED_TWO;
  BY &UNIQ_ID OUTC_TYPE T;
RUN;

%LET PARMS = ALPHA0=&SV_INTERCEPT;

%DO I=1 %TO &NUMB_RECR;
  %IF (&I = 1) %THEN %DO;
   %LET OUTSTR_BHR = R&I * EVENT_R&I;
   %LET OUTSTR_CBHR = R&I * DUR_R&I;
   %LET OUTSTR_R = R&I;
   %LET PARMS = &PARMS R&I=%SCAN(&SV_RECR,&I,' ');
  %END;
  %ELSE %DO;
   %LET OUTSTR_BHR = &OUTSTR_BHR + R&I * EVENT_R&I;
   %LET OUTSTR_CBHR = &OUTSTR_CBHR + R&I * DUR_R&I;
   %LET OUTSTR_R = &OUTSTR_R R&I;
   %LET PARMS = &PARMS R&I=%SCAN(&SV_RECR,&I,' ');
```

```
 %END;
%END;

%IF (&NUMB_COV_RECR = 0) %THEN %DO;
 %LET OUTSTR_MU1 = U;
%END;
%ELSE %DO;
 %DO I=1 %TO &NUMB_COV_RECR;
  %IF (&I = 1) %THEN %DO;
   %LET OUTSTR_MU1 = BETA&I * %SCAN(&COV_RECR,&I);
   %LET PARMS = &PARMS BETA&I=%SCAN(&SV_BETA,&I,' ');
  %END;
  %ELSE %DO;
   %LET OUTSTR_MU1 = &OUTSTR_MU1 + BETA&I * %SCAN(&COV_RECR,&I);
   %LET PARMS = &PARMS BETA&I=%SCAN(&SV_BETA,&I,' ');
  %END;
  %IF (&I = &NUMB_COV_RECR) %THEN %DO;
   %LET OUTSTR_MU1 = &OUTSTR_MU1 + U;
  %END;
 %END;
%END;

%IF (&NUMB_COV_LONG = 0) %THEN %DO;
 %LET OUTSTR_MU3 = ALPHA0 + ALPHA1*T;
 %LET PARMS = &PARMS ALPHA1=%SCAN(&SV_TIME,1,' ');
 %DO K=1 %TO &NUMB_KNTS;
  %LET OUTSTR_MU3 = &OUTSTR_MU3 + ALPHA_S0_&K * TIME_SPLINE_&K ;
```

```
   %LET PARMS = &PARMS ALPHA_S0_&K=%SCAN(&SV_TIME,&K+1,' ');
  %END;
%END;
%ELSE %DO;
  %DO I=1 %TO &NUMB_COV_LONG;
   %IF (&I = 1) %THEN %DO;
    %LET OUTSTR_MU3 = ALPHA0 + ALPHA1*T;
    %LET PARMS = &PARMS ALPHA1=%SCAN(&SV_TIME,1,' ');
    %DO K=1 %TO &NUMB_KNTS;
     %LET OUTSTR_MU3 = &OUTSTR_MU3 + ALPHA_S0_&K * TIME_SPLINE_&K ;
          %LET PARMS = &PARMS ALPHA_S0_&K=%SCAN(&SV_TIME,&K+1,' ');
    %END;
    %LET OUTSTR_MU3 = &OUTSTR_MU3 + ALPHA_I&I * %SCAN(&COV_LONG,&I)
        ;
    %LET PARMS = &PARMS ALPHA_I&I=%SCAN(&SV_COV_LONG,&I,' ');
   %END;
   %ELSE %DO;
    %LET OUTSTR_MU3 = &OUTSTR_MU3 + ALPHA_I&I * %SCAN(&COV_LONG,&I)
        ;
    %LET PARMS = &PARMS ALPHA_I&I=%SCAN(&SV_COV_LONG,&I,' ');
   %END;
  %END;
%END;

%IF (&NUMB_COV_LONG_TIME NE 0) %THEN %DO;
  %DO J=1 %TO &NUMB_COV_LONG_TIME;
   %LET OUTSTR_MU3 = &OUTSTR_MU3 + ALPHA_S&J * T * %SCAN(&
```

```
      COV_LONG_TIME,&J);
   %LET PARMS = &PARMS ALPHA_S&J=%SCAN(&SV_COV_LONG_TIME,1,' ');
   %DO K=1 %TO &NUMB_KNTS;
    %LET OUTSTR_MU3 = &OUTSTR_MU3 + ALPHA_S&J._&K * TIME_SPLINE_&K
       * %SCAN(&COV_LONG_TIME,&J);
    %LET PARMS = &PARMS ALPHA_S&J._&K=%SCAN(&SV_COV_LONG_TIME,&K
       +1,' ');
   %END;
  %END;
%END;

%LET PARMS = &PARMS GAMMA1=%SCAN(&SV_GAMMA,1,' ');
%LET PARMS = &PARMS VARU=%SCAN(&SV_VAR,1,' ') VARV=%SCAN(&SV_VAR
    ,2,' ') VARE=%SCAN(&SV_VAR,3,' ');
%LET OUTSTR_MU3 = &OUTSTR_MU3 + GAMMA1 * U + V;

PROC NLMIXED DATA=NLMIXED_TWO QPOINTS=5 MINITER=&MINITER;
 PARMS &PARMS;
 BOUNDS &OUTSTR_R VARU VARV VARE >= 0;

 IF OUTC_TYPE = 1 THEN DO;
  MU3 = &OUTSTR_MU3;
  LOGLIK = -.5*(Y - MU3)**2/(VARE) -.5*LOG(2*3.14159*VARE);
 END;

 ELSE IF OUTC_TYPE = 2 THEN DO;
  BASE_HAZ_R = &OUTSTR_BHR;
```

```
    CUM_BASE_HAZ_R = &OUTSTR_CBHR;

    MU1 = &OUTSTR_MU1;
    LOGLIK1 = -EXP(MU1)*CUM_BASE_HAZ_R;

    IF LAST = 0 THEN LOGLIK = LOG(BASE_HAZ_R) + MU1;
    ELSE IF LAST = 1 THEN LOGLIK = LOG(BASE_HAZ_R) + MU1 + LOGLIK1;

  END;

  MODEL OUTC ~ GENERAL(LOGLIK);
  RANDOM U V ~ NORMAL([0,0],[VARU,0,VARV]) SUBJECT=&UNIQ_ID OUT=
      EM_BAYES_EST;

  ODS OUTPUT PARAMETERESTIMATES=PE;
  PREDICT BASE_HAZ_R*EXP(MU1) OUT=FULL_HAZARD;

RUN;

/*** USING THE TWO-PART MODEL OUTPUT TO DETERMINE INDEPENDENT/
    DEPENDENT TERMINATION ***/

DATA DAT_LAST;
  SET DAT1;
  IF LAST=1 THEN OUTPUT;
RUN;
DATA EM_BAYES_NEW;
```

```
  SET EM_BAYES_EST;
  WHERE EFFECT='U';
  U_BAYES_EST = ESTIMATE;
  U_BAYES_PRED_STD = STDERRPRED;
  U_LOWER = LOWER;
  U_UPPER = UPPER;
  DROP ESTIMATE STDERRPRED LOWER UPPER DF TVALUE PROBT ALPHA;
RUN;
PROC SORT DATA=EM_BAYES_NEW; BY &UNIQ_ID; RUN;
PROC SORT DATA=DAT_LAST; BY &UNIQ_ID; RUN;
DATA DAT_PRED;
  MERGE DAT_LAST (IN=A) EM_BAYES_NEW (IN=B);
  BY &UNIQ_ID;
  IF A=1;
RUN;
DATA DAT_PRED;
  RETAIN R_:;
  IF _N_ = 1 THEN SET PCTL_RECR;
  SET DAT_PRED;
  AA=1;
RUN;
PROC TRANSPOSE DATA=PE OUT=PE_TRANS;
  ID PARAMETER;
  VAR ESTIMATE;
RUN;
DATA PE_TRANS; SET PE_TRANS; AA=1; RUN;
DATA DAT_PRED (DROP=AA _NAME_);
```

```
 MERGE DAT_PRED PE_TRANS;
 BY AA;
RUN;

%MACRO RANDEXP(SIGMA);
          ((&SIGMA)*RAND("Exponential"))
%MEND;

%LET NUMB_COV_RECR = %SYSFUNC(COUNTW(%STR(&COV_RECR)));
%IF (&NUMB_COV_RECR = 0) %THEN %DO;
 %LET OUTSTR_MU1 = U_BAYES_EST;
%END;
%ELSE %DO;
 %DO I=1 %TO &NUMB_COV_RECR;
  %IF (&I = 1) %THEN %DO;
   %LET OUTSTR_MU1 = BETA&I * %SCAN(&COV_RECR,&I);
  %END;
  %ELSE %DO;
   %LET OUTSTR_MU1 = &OUTSTR_MU1 + BETA&I * %SCAN(&COV_RECR,&I);
  %END;
  %IF (&I = &NUMB_COV_RECR) %THEN %DO;
   %LET OUTSTR_MU1 = &OUTSTR_MU1 + U_BAYES_EST;
  %END;
 %END;
%END;

DATA DAT_PRED2;
```

```
  RETAIN R_:;
  SET DAT_PRED;
  ARRAY QUANT_R {*} R_:;
  ARRAY R{&NUMB_RECR};
  ARRAY HAZARD {&NUMB_RECR};
  DO I=1 TO &NUMB_RECR;
    IF QUANT_R{I} < T <= QUANT_R{I+1} THEN HAZARD_INTERVAL = I;
    HAZARD{I} = R{I}*EXP(&OUTSTR_MU1);
  END;
  DROP I;
RUN;

DATA DAT_PRED3;
  RETAIN R_:;
  SET DAT_PRED2;
  ARRAY QUANT_R {*} R_:;
  ARRAY S{&NUMB_RECR};
  ARRAY F{&NUMB_RECR};
  ARRAY LIM{&NUMB_RECR};
  ARRAY HAZARD {&NUMB_RECR};

  CALL STREAMINIT(&SEED);
  X=%RANDEXP(1);

  DO I=1 TO &NUMB_RECR;
    IF HAZARD_INTERVAL = I THEN DO;
      DO J=1 TO &NUMB_RECR;
```

```
      IF  (I+J-1) <= &NUMB_RECR THEN DO;
       S{J} = QUANT_R{I+J}-T;
            F{J} = HAZARD{I+J-1};
           END;
           ELSE DO;
       S{J} = 0;
            F{J} = 0;
           END;
    END;
  END;
 END;
 DO I=1 TO &NUMB_RECR;
  LIM{I} = 0;
  IF I > 1 THEN DO;
   DO J=1 TO (I-1);
    IF J=1 THEN LIM{I} = LIM{I} + F{J}*S{J};
         ELSE IF J > 1 THEN LIM{I} = LIM{I} +F{J}*(S{J}-S{J-1});
   END;
  END;
 END;
 DROP I J;
RUN;

DATA DAT_PRED4;
 SET DAT_PRED3;
 ARRAY S{&NUMB_RECR};
 ARRAY F{&NUMB_RECR};
```

```
ARRAY LIM{&NUMB_RECR};

DO I=1 TO &NUMB_RECR;
 IF HAZARD_INTERVAL = I THEN DO;
  DO J=1 TO (&NUMB_RECR-I+1);
   IF J = 1 & J NE (&NUMB_RECR-I+1) THEN DO;
        IF LIM{J} <= X <= LIM{J+1} THEN EXPECTED_TIME = (X-LIM{J
          })/F{J};
       END;
       ELSE IF J = 1 & J = (&NUMB_RECR-I+1) THEN DO;
        IF LIM{J} <= X THEN EXPECTED_TIME = (X-LIM{J})/F{J};
       END;
   ELSE IF J NE (&NUMB_RECR-I+1) THEN DO;
        IF LIM{J} < X <= LIM{J+1} THEN EXPECTED_TIME = S{J-1} + (
          X-LIM{J})/F{J};
       END;
       ELSE DO;
        IF LIM{J} < X THEN EXPECTED_TIME = S{J-1} + (X-LIM{J})/F{
          J};
       END;
  END;
 END;
END;

EXPECTED_NEXT_VISIT_TIME = SUM(OF EXPECTED_TIME T);
EXPECTED_NEXT_VISIT_DATE = SUM(OF EXPECTED_TIME &TIME);
```

```
 DROP I J;
RUN;

PROC SORT DATA=DAT1; BY &UNIQ_ID T; RUN;
PROC SORT DATA=DAT_PRED4; BY &UNIQ_ID; RUN;
DATA DAT_NEXT;
 FORMAT  EXPECTED_NEXT_VISIT_DATE MMDDYY10.;
 MERGE DAT1 (IN=IND1)
       DAT_PRED4 (IN=IND2 KEEP=&UNIQ_ID
                          EXPECTED_NEXT_VISIT_TIME
                          EXPECTED_NEXT_VISIT_DATE);
 IF EXPECTED_NEXT_VISIT_TIME GE CI THEN DELTAI = 0; ELSE DELTAI =
    1;
 BY &UNIQ_ID;
 IF IND1=1;
 IF LAST NE 1 THEN CENS = 0;
 ELSE IF LAST = 1 & DELTAI = 1 THEN CENS = 2;
 ELSE IF LAST = 1 & DELTAI = 0 THEN CENS = 1;
RUN;

/*** FITTING THE THREE-PART MODEL ***/

DATA TERM;
 SET DAT_NEXT;
 IF CENS = 2;
RUN;
PROC UNIVARIATE DATA=TERM NOPRINT;
```

```
 VAR DI;
 OUTPUT PCTLPRE=T_ PCTLPTS=0 TO 99 BY &INCREMENT_TERM,100  OUT=
     PCTL_TERM;
RUN;
DATA PCTL;
 MERGE PCTL_RECR PCTL_TERM;
RUN;

DATA DAT_NEXT1;
 SET LONG (IN=IND1)
      DAT_NEXT (IN=IND2)
      DAT_NEXT (WHERE=(LAST=1) IN=IND3);
 IF IND1 THEN DO;
   OUTC = Y;
   OUTC_TYPE = 1;
 END;
 ELSE IF IND2 THEN DO;
  OUTC = T;
  OUTC_TYPE = 2;
  CENS = 0;
 END;
 ELSE IF IND3 THEN DO;
  IF DELTAI = 1 THEN OUTC = DI;
  ELSE IF DELTAI = 0 THEN OUTC = CI;
  OUTC_TYPE = 2;
 END;
 IF OUTC NE 0;
```

```
RUN;

DATA NLMIXED_DAT;
 IF _N_ = 1 THEN SET PCTL;
 SET DAT_NEXT1;

 ARRAY QUANT_R {*} R_:;
 ARRAY QUANT_T {*} T_:;

 ARRAY DUR_R {&NUMB_RECR};
 ARRAY DUR_T {&NUMB_TERM};

 ARRAY EVENT_R {&NUMB_RECR};
 ARRAY EVENT_T {&NUMB_TERM};

 DO I=1 TO &NUMB_RECR;
  DUR_R{I} = 0;
  EVENT_R{I} = 0;
 END;

 DO I=1 TO &NUMB_TERM;
  DUR_T{I} = 0;
  EVENT_T{I} = 0;
 END;

 IF OUTC_TYPE = 2 & CENS = 0 THEN DO;
  DO I=2 TO (&NUMB_RECR + 1);
```

```
    IF OUTC <= QUANT_R{I} THEN DO;
     EVENT_R{I-1} = 1;
     I = (&NUMB_RECR + 1);
    END;
  END;
END;

ELSE IF OUTC_TYPE = 2 THEN DO;
  DO I=2 TO (&NUMB_RECR + 1);
    IF OUTC <= QUANT_R{I} THEN DO;
     DUR_R{I-1} = MAX(OUTC-QUANT_R{I-1},0);
     I = (&NUMB_RECR + 1);
    END;
    ELSE DUR_R{I-1} = QUANT_R{I}-QUANT_R{I-1};
  END;

  DO I=2 TO (&NUMB_TERM + 1);
    IF OUTC <= QUANT_T{I} THEN DO;
     EVENT_T{I-1}= (CENS=2);
     DUR_T{I-1} = MAX(OUTC-QUANT_T{I-1},0);
     I = (&NUMB_TERM + 1);
    END;
    ELSE DUR_T{I-1} = QUANT_T{I}-QUANT_T{I-1};
  END;
END;

DROP I;
```

```
RUN;

PROC SORT DATA=NLMIXED_DAT; BY &UNIQ_ID OUTC_TYPE T; RUN;

%LET PARMS = ALPHA0=&SV_INTERCEPT;

%DO I=1 %TO &NUMB_RECR;
 %IF (&I = 1) %THEN %DO;
  %LET OUTSTR_BHR = R&I * EVENT_R&I;
  %LET OUTSTR_CBHR = R&I * DUR_R&I;
  %LET OUTSTR_R = R&I;
  %LET PARMS = &PARMS R&I=%SCAN(&SV_RECR,&I,' ');
 %END;
 %ELSE %DO;
  %LET OUTSTR_BHR = &OUTSTR_BHR + R&I * EVENT_R&I;
  %LET OUTSTR_CBHR = &OUTSTR_CBHR + R&I * DUR_R&I;
  %LET OUTSTR_R = &OUTSTR_R R&I;
  %LET PARMS = &PARMS R&I=%SCAN(&SV_RECR,&I,' ');
 %END;
%END;

%DO I=1 %TO &NUMB_TERM;
 %IF (&I = 1) %THEN %DO;
  %LET OUTSTR_BHT = T&I * EVENT_T&I;
  %LET OUTSTR_CBHT = T&I * DUR_T&I;
  %LET OUTSTR_T = T&I;
  %LET PARMS = &PARMS T&I=%SCAN(&SV_TERM,&I,' ');
```

```
 %END;
 %ELSE %DO;
  %LET OUTSTR_BHT = &OUTSTR_BHT + T&I * EVENT_T&I;
  %LET OUTSTR_CBHT = &OUTSTR_CBHT + T&I * DUR_T&I;
  %LET OUTSTR_T = &OUTSTR_T T&I;
  %LET PARMS = &PARMS T&I=%SCAN(&SV_TERM,&I,' ');
 %END;
%END;

%IF (&NUMB_COV_RECR = 0) %THEN %DO;
 %LET OUTSTR_MU1 = U;
%END;
%ELSE %DO;
 %DO I=1 %TO &NUMB_COV_RECR;
  %IF (&I = 1) %THEN %DO;
   %LET OUTSTR_MU1 = BETA&I * %SCAN(&COV_RECR,&I);
   %LET PARMS = &PARMS BETA&I=%SCAN(&SV_BETA,&I,' ');
  %END;
  %ELSE %DO;
   %LET OUTSTR_MU1 = &OUTSTR_MU1 + BETA&I * %SCAN(&COV_RECR,&I);
   %LET PARMS = &PARMS BETA&I=%SCAN(&SV_BETA,&I,' ');
  %END;
  %IF (&I = &NUMB_COV_RECR) %THEN %DO;
   %LET OUTSTR_MU1 = &OUTSTR_MU1 + U;
  %END;
 %END;
%END;
```

```
%IF (&NUMB_COV_TERM = 0) %THEN %DO;
 OUTSTR_MU2 = GAMMA2 * U + GAMMA3 * V;
%END;
%ELSE %DO;
 %DO I=1 %TO &NUMB_COV_TERM;
  %IF (&I = 1) %THEN %DO;
   %LET OUTSTR_MU2 = ETA&I * %SCAN(&COV_TERM,&I);
   %LET PARMS = &PARMS ETA&I=%SCAN(&SV_ETA,&I,' ');
  %END;
  %ELSE %DO;
   %LET OUTSTR_MU2 = &OUTSTR_MU2 + ETA&I * %SCAN(&COV_TERM,&I);
   %LET PARMS = &PARMS ETA&I=%SCAN(&SV_ETA,&I,' ');
  %END;
  %IF (&I = &NUMB_COV_TERM) %THEN %DO;
   %LET OUTSTR_MU2 = &OUTSTR_MU2 + GAMMA2 * U + GAMMA3 * V;
  %END;
 %END;
%END;

%IF (&NUMB_COV_LONG = 0) %THEN %DO;
 %LET OUTSTR_MU3 = ALPHA0 + ALPHA1*T;
 %LET PARMS = &PARMS ALPHA1=%SCAN(&SV_TIME,1,' ');
 %DO K=1 %TO &NUMB_KNTS;
  %LET OUTSTR_MU3 = &OUTSTR_MU3 + ALPHA_S0_&K * TIME_SPLINE_&K ;
  %LET PARMS = &PARMS ALPHA_S0_&K=%SCAN(&SV_TIME,&K+1,' ');
 %END;
```

```
%END;
%ELSE %DO;
 %DO I=1 %TO &NUMB_COV_LONG;
  %IF (&I = 1) %THEN %DO;
   %LET OUTSTR_MU3 = ALPHA0 + ALPHA1*T;
   %LET PARMS = &PARMS ALPHA1=%SCAN(&SV_TIME,1,' ');
   %DO K=1 %TO &NUMB_KNTS;
    %LET OUTSTR_MU3 = &OUTSTR_MU3 + ALPHA_S0_&K * TIME_SPLINE_&K ;
        %LET PARMS = &PARMS ALPHA_S0_&K=%SCAN(&SV_TIME,&K+1,' ');
   %END;
   %LET OUTSTR_MU3 = &OUTSTR_MU3 + ALPHA_I&I * %SCAN(&COV_LONG,&I)
       ;
   %LET PARMS = &PARMS ALPHA_I&I=%SCAN(&SV_COV_LONG,&I,' ');
  %END;
  %ELSE %DO;
   %LET OUTSTR_MU3 = &OUTSTR_MU3 + ALPHA_I&I * %SCAN(&COV_LONG,&I)
       ;
   %LET PARMS = &PARMS ALPHA_I&I=%SCAN(&SV_COV_LONG,&I,' ');
  %END;
 %END;
%END;

%IF (&NUMB_COV_LONG_TIME NE 0) %THEN %DO;
 %DO J=1 %TO &NUMB_COV_LONG_TIME;
  %LET OUTSTR_MU3 = &OUTSTR_MU3 + ALPHA_S&J * T * %SCAN(&
      COV_LONG_TIME,&J);
  %LET PARMS = &PARMS ALPHA_S&J=%SCAN(&SV_COV_LONG_TIME,1,' ');
```

```
   %DO K=1 %TO &NUMB_KNTS;
    %LET OUTSTR_MU3 = &OUTSTR_MU3 + ALPHA_S&J._&K * TIME_SPLINE_&K
        * %SCAN(&COV_LONG_TIME,&J);
    %LET PARMS = &PARMS ALPHA_S&J._&K=%SCAN(&SV_COV_LONG_TIME,&K
        +1,' ');
   %END;
  %END;
%END;

%LET PARMS = &PARMS GAMMA1=%SCAN(&SV_GAMMA,1,' ') GAMMA2=%SCAN(&
    SV_GAMMA,2,' ') GAMMA3=%SCAN(&SV_GAMMA,1,' ');
%LET PARMS = &PARMS VARU=%SCAN(&SV_VAR,1,' ') VARV=%SCAN(&SV_VAR
    ,2,' ') VARE=%SCAN(&SV_VAR,3,' ');
%LET OUTSTR_MU3 = &OUTSTR_MU3 + GAMMA1 * U + V;

PROC NLMIXED DATA=NLMIXED_DAT QPOINTS=5 MINITER=&MINITER;
 PARMS &PARMS;
 BOUNDS &OUTSTR_R &OUTSTR_T VARU VARV VARE >= 0;

 IF OUTC_TYPE = 1 THEN DO;
  MU3 = &OUTSTR_MU3;
  LOGLIK = -.5*(Y - MU3)**2/(VARE) -.5*LOG(2*3.14159*VARE);
 END;

 ELSE IF OUTC_TYPE = 2 THEN DO;
  BASE_HAZ_R = &OUTSTR_BHR;
  CUM_BASE_HAZ_R = &OUTSTR_CBHR;
```

```
  BASE_HAZ_T = &OUTSTR_BHT;
  CUM_BASE_HAZ_T = &OUTSTR_CBHT;

  MU1 = &OUTSTR_MU1;
  MU2 = &OUTSTR_MU2;
  LOGLIK1 = -EXP(MU1)*CUM_BASE_HAZ_R;
  LOGLIK2 = -EXP(MU2)*CUM_BASE_HAZ_T;

  IF CENS = 0 THEN LOGLIK = LOG(BASE_HAZ_R) + MU1;
  ELSE IF CENS = 2 THEN LOGLIK = LOGLIK1 + LOGLIK2 + LOG(
      BASE_HAZ_T) + MU2;
  ELSE IF CENS = 1 THEN LOGLIK = LOGLIK1 + LOGLIK2;
 END;

 MODEL OUTC ~ GENERAL(LOGLIK);
 RANDOM U V ~ NORMAL([0,0],[VARU,0,VARV]) SUBJECT=&UNIQ_ID OUT=
     EM_BAYES_EST_FINAL;

 ODS OUTPUT PARAMETERESTIMATES=PE_FINAL;

RUN;

%PUT LONGITUDINAL PARAMETER FUNCTION: &OUTSTR_MU3;
%PUT RECURRENT EVENT PARAMETER FUNCTION: &OUTSTR_MU1;
%PUT TERMINAL EVENT PARAMETER FUNCTION: &OUTSTR_MU2;
%PUT BASELINE HAZARD FOR RECURRENT EVENTS: &OUTSTR_BHR;
```

```
%PUT CUMULATIVE BASELINE HAZARD FOR RECURRENT EVENTS: &OUTSTR_CBHR
    ;
%PUT BASELINE HAZARD FOR TERMINAL EVENTS: &OUTSTR_BHR;
%PUT CUMULATIVE BASELINE HAZARD FOR TERMINAL EVENTS: &OUTSTR_CBHR;

%MEND;

/**** EXAMPLE OF IMPLEMENTING THE MACRO ON SIMULATED DATA ***/

*%SHARED3_RANEFF(DAT=SIMDAT_1, Y=Y, TIME=DAY, UNIQ_ID=PID,
   TIME_END=360.64, TIME_TERM=TERM,
                    COV_LONG = Z,
                                COV_RECR = Z,
                                COV_TERM = Z,
                                NUMB_RECR = 10,
                                NUMB_TERM = 10,
                    DURATION_END = 360.64,
                                MINITER=250,
                    SV_RECR = 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01
                       0.01 0.01,
                    SV_TERM = 0.001 0.001 0.002 0.003 0.003 0.004
                       0.006 0.004 0.008 0.012,
                    SV_INTERCEPT = 10,
                    SV_TIME = 0.2,
                    SV_COV_LONG = 1,
                    SV_BETA = 1,
                    SV_ETA = 1,
```

```
                    SV_GAMMA = 1.5  -0.5  1,
                    SV_VAR = 1  0.5  1);

/** PARAMETER ESTIMATES CONTAINED WITHIN DATA SET WORK.PE_FINAL
    **/
/** INDIVIDUAL-LEVEL EMPIRICAL BAYES ESTIMATES FOR UI CONTAINED
    WITHIN WORK.EM_BAYES_EST_FINAL **/
/** MACRO WILL ALSO OUTPUT PARAMETER FUNCTION SPECIFICATIONS TO
    THE LOG AFTER THE FINAL PROC NLMIXED TO ASSIST WITH
    INTERPRETATION OF PARAMETER ESTIMATE NAMES **/


/** INTERPRETATION OF PARAMETERS **/

*ALPHA0 = INTERCEPT;
*R.. = BASELINE HAZARD ESTIMATES FOR THE RECURRECT EVENTS
    COMPONENT OF THE MODEL (NUMBER OF ESTIMATES = NUMB_RECR);
*T.. = BASELINE HAZARD ESTIMATES FOR THE TERMINAL EVENTS COMPONENT
     OF THE MODEL (NUMBER OF ESTIMATES = NUMB_TERM);
*BETA.. = COEFFICIENTS ASSOCIATED WITH COVARIATES INCLUDED IN THE
    RECURRENT EVENTS COMPONENT OF THE MODEL (NUMBER OF ESTIMATES =
    LENGTH OF COV_RECR);
*ETA.. = COEFFICIENTS ASSOCIATED WITH COVARIATES INCLUDED IN THE
    TERMINANL EVENTS COMPONENT OF THE MODEL (NUMBER OF ESTIMATES =
    LENGTH OF COV_TERM);
*ALPHA1 = COEFFICIENT ASSOCIATED WITH TIME IN THE LONGITUDINAL
    COMPONENT OF THE MODEL;
```

*ALPHA_S0_.. = COEFFICIENTS ASSOCIATED WITH EACH OF THE SPLINE KNOTS (NUMBER OF ESTIMATES = LENGTH OF KNTS_SPLINE);

*ALPHA_S.. = COEFFICIENTS ASSOCIATED WITH SLOPE TERMS (INTERACTED WITH TIME) IN THE LONGITUDINAL COMPONENT OF THE MODEL (NUMBER OF ESTIMATES = LENGTH OF COV_LONG_TIME);

*ALPHA_S.._... = COEFFICIENTS ASSOCIATED WITH SLOPE TERMS (INTERACTED WITH EACH OF THE SPLINE KNOTS) IN THE LONGITUDINAL COMPONENT OF THE MODEL (NUMBER OF ESTIMATES = LENGTH OF COV_LONG_TIME MULTIPLIED BY LENGTH OF KNTS_SPLINE);

*ALPHA_I.. = COEFFICIENTS ASSOCIATED WITH COVARIATES INCLUDED IN THE LONGITUDINAL COMPONENT OF THE MODEL (NUMBER OF ESTIMATES = LENGTH OF COV_LONG);

*GAMMA1 = COEFFICIENT ON THE RANDOM EFFECT UI IN THE LONGITUDINAL COMPONENT OF THE MODEL;

*GAMMA2 = COEFFICIENT ON THE RANDOM EFFECT UI IN THE TERMINAL EVENTS COMPONENT OF THE MODEL;

*GAMMA3 = COEFFICIENT ON THE RANDOM EFFECT VI IN THE TERMINAL EVENTS COMPONENT OF THE MODEL;

*VARU = VARIANCE OF THE RANDOM EFFECT UI SHARED ACROSS ALL THREE COMPONENTS OF THE MODEL;

*VARV = VARIANCE OF THE RANDOM EFFECT VI SHARED ACROSS THE LONGITUINAL AND TERMINAL EVENTS COMPONENTS OF THE MODEL;

*VARE = VARIANCE OF THE RANDOM EFFECT EIJ PRESENT IN THE LONGITUDINAL COMPONENT OF THE MODEL;

REFERENCES

[1] J. Jeffrey, A. Klomhaus, H. Aralis, W. Barrera, S. Rosenberg, M. Grossman, and P. Lester. Factors associated with response and remission from depression at 6-months of treatment in a retrospective cohort treated within an integrated care program. *BMC Health Services Research*, 21(703), 2021.

[2] K.L. Myrick and D.F. Ogburn. Percentage of office-based physicians using any electronic health record (EHR)/electronic medical record (EMR) system and physicians that have a certified EHR/EMR system, by selected characteristics: National Electronic Health Records Survey, 2017. *National Center for Health Statistics*, 2019. Retrieved from, `https://www.cdc.gov/nchs/data/nehrs/2017_NEHRS_web_table_EHR_Variation.pdf`.

[3] E. Kim, S.M. Rubinstein, K.T. Nead, A.P. Wojcieszynski, P.E. Gabriel, and J.L. Warner. The Evolving Use of Electronic Health Records (EHR) for Research. *Seminars in Radiation Oncology*, 29:354–361, 2019.

[4] W.R. Hersh, M.G. Weiner, P.J. Embi, J.R. Logan, P.R.O. Payne, E.V. Bernstam, H.P. Lehmann, G. Hripcsak, T.H. Hartzog, J.J. Cimino, and J.H. Saltz. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Med Care*, 51(8 Suppl 3):S30–S37, 2013.

[5] F.M. Blyth, L.M. March, A.J.M. Brnabic, and M.J. Cousins. Chronic pain and frequent use of health care. *Pain*, 111(1-2):S30–S37, 2004.

[6] J. Oosterhaven, H. Wittink, J. Mollema, C. Kruitwagen, and W. Deville. Predictors of dropout in interdisciplinary chronic chronic pain management programmes: a systematic review. *J Rehabil Med*, 51:2–10, 2019.

[7] J.G. Hanly, K. Thompson, and C. Skedgel. Utilization of Ambulatory Physician Encounters, Emergency Room Visits, and Hospitalizations by Systematic Lupus Erythematosus Patients: A 13-Year Population Health Study. *Arthritis Care Res*, 68(8):1128–1134, 2016.

[8] US Department of Health and Human Services, Substance Abuse and Mental Health Services Administration. Samhsa – behavioral health integration. Retrieved from, `https://www.samhsa.gov/`.

[9] D.J. Roseborough, J.T. McLeod, and F.I. Wright. Attrition in Psychotherapy: A Survival Analysis. *Res Soc Work Pract"*, 26(7):803–815, 2016.

[10] K. Possemato, M. Wade, J. Andersen, and P. Ouimette. Impact of PTSD, Depression, and Substance Use Disorders on Disease Burden and Health Care Utilization Among OEF/OIF Veterans. *Psychol Trauma*, 2(3):218–223, 2010.

[11] K.I. Howard, S.M. Kopta, M.S. Krause, and D.E. Orlinsky. The Dose-Effect Relationship in Psychotherapy. *American Psychologist*, 41(2):159–164, 1986.

[12] R.J. Cook and J.F. Lawless. Analysis of repeated events. *Statistical Methods in Medical Research*, 11:141–166, 2002.

[13] L. Liu, R.A. Wolfe, and X. Huang. Shared Frailty Models for Recurrent Events and a Terminal Event. *Biometrics*, 60:747–756, 2004.

[14] B.A. Goldstein, M. Phelan, N.J. Pagidipati, and S.B. Peskoe. How and when informative visit processes can bias inference when using electronic health records data for clinical research. *J Am Med Inform Assoc*, 26(12):1609–1617, 2019.

[15] L.-A. Lin, S. Luo, B.E. Chen, and B.R. Davis. Bayesian analysis of multi-type recurrent events and dependent termination with nonparametric covariate functions. *Stat Methods Med Res*, 26(6):2869–2884, 2017.

[16] L. Liu, X. Huang, and J. O'Quigley. Analysis of Longitudinal Data in the Presence of Informative Observational Times and a Dependent Terminal Event. *Biometrics*, 64:750–758, 2008.

[17] Y. Liang, W. Lu, and Z. Ying. Joint Modeling and Analysis of Longitudinal Data with Informative Observation Times. *Biometrics*, 65:377–384, 2009.

[18] N.G. Weiskopf, A. Rusanov, and C. Weng. Sick patients have more data: the non-random completeness of electronic health records. *AMIA Annu Symp Proc 2013*, pages 1472–1477, November 16, 2013.

[19] R. Westmacott and J. Hunsley. Reasons for Terminating Psychotherapy: A General Population Study. *Journal of Clinical Psychology*, 66(9):965–977, 2010.

[20] J. Wang. Mental Health Treatment Dropout and Its Correlates in a General Population Sample. *Medical Care*, 45(3):224–229, 2007.

[21] Substance Abuse and Mental Health Services Administration. Key substance use and mental health indicators in the United States: Results from the 2017 National Survey on Drug Use and Health (HHS Publication No. SMA 18-5068, NSDUH Series H-53), 2018.

[22] T. Vos, R.M. Barber, B. Bell, A. Bertozzi-Villa, S. Biryukov, I. Bolliger, F. Charlson, A. Davis, L. Degenhardt, D. Dicker, and L. Duan. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013. *The Lancet*, 386(9995):743–800, 2013.

[23] K. Kroenke and J. Unutzer. Closing the False Divide: Sustainable Approaches to Integrating Mental Health Services into Primary Care. *Journal of General Internal Medicine*, 32(4):404–410, 2017.

[24] P.J. Cunningham. Beyond Parity: Primary Care Physicians' Perspectives On Access To Mental Health Care. *Health Affairs*, 28(3):490–501, 2009.

[25] C.L. Barry and H.A. Huskamp. Moving beyond Parity - Mental Health and Addiction Care under the ACA. *The New England Journal of Medicine*, 365(11):973–975, 2011.

[26] N.M.A. Faghri, C.M. Boisvert, and S. Faghri. Understanding the expanding role of primary care physicians (PCPs) to primary psychiatric care physicians (PPCPs): enhancing the assessment and treatment of psychiatric conditions. *Mental Health in Family Medicine*, 7:17–25, 2010.

[27] J. Archer, P. Bower, S. Gilbody, K. Lovell, D. Richards, L. Gask, C. Dickens, and P. Coventry. Collaborate care for depression and anxiety problems. *Cochrane Database of Systematic Reviews*, 10, 2012.

[28] E.W.J. Kokkonen, S.A. Davis, H-C. Lin, T.S. Dabade, S.R. Feldman, and A.B. Fleischer Jr. Use of electronic medical records differs by specialty and office settings. *Journal of the American Medical Informatics Association*, 20:e33–e38, 2013.

[29] R.M.A. Clarke, J. Jeffrey, M. Grossman, T. Strouse, M. Gitlin, and S.A. Skootsky. Delivering On Accountable Care: Lessons From A Behavioral Health Program To Improve Access And Outcomes. *Health Affairs*, 35(8):1487–1493, 2016.

[30] K. Kroenke, R.L. Spitzer, and J.B.W. Williams. The PHQ–9. *J Gen Intern Med*, 16:606–613, 2001.

[31] R.L. Spitzer, K. Kroenke, J.B. Williams, and B. Löwe. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med*, 166(10):1092–1097, 2006.

[32] A. Prins, P. Ouimette, R. Kimerling, R.P. Cameron, D.S. Hugelshofer, J. Shaw-Hegwer, A. Thrailkill, F.D. Gusman, and J.I. Sheikh. The Primary Care PTSD Screen (PC-PTSD): Development and operating characteristics. *Primary Care Psychiatry*, 9:9–14, 2003.

[33] E.A. Walker, E. Newman, D.J. Dobie, P. Ciechanowski, and W. Katon. Validation of the PTSD checklist in an HMO sample of women. *Gen Hosp Psychiatry*, 24(6):375–380, 2002.

[34] K. Bush, D.R. Kivlahan, M.B. McDonell, S.D. Fihn, and K.A. Bradley. The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. *Arch Intern Med*, 158(16):1789–1795, 1998.

[35] E. Yudko, O. Lozhkina, and A. Fouts. A comprehensive review of the psychometric properties of the Drug Abuse Screening Test. *J Subst Abuse Treat*, 32(2):189–198, 2007.

[36] C. Faucett and D.C. Thomas. Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates: A Gibbs Sampling Approach. *Stat Med*, 15:1663–1685, 1996.

[37] L. Liu, R.A. Wolfe, and J.D Kalbfleisch. A shared random effects model for censored medical costs and mortality. *Stat Med*, 26(1):139–155, 2007.

[38] T. Lancaster and O. Intrator. Panel Data with Survival: Hospitalization of HIV-Positive Patients. *J Am Stat Assoc*, 93(441):46–53, 1998.

[39] X. Guo and B.P. Carlin. Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages. *Am Stat*, 58(1):16–24, 2004.

[40] L. Liu and X. Huang. The use of Gaussian quadrature for estimation in frailty proportional hazards models. *Stat Med*, 27:2665–2683, 2008.

[41] L. Liu and X. Huang. Joint analysis of correlated repeated measures and recurrent events processes in the presence of death with application to a study on acquired immune deficiency syndrome. *Appl Statist*, 58(1):65–81, 2009.

[42] K.-Y. Liu and X. Zhao. Robust estimation for longitudinal data with informative observation times. *Can J Stat*, 43(4):519–533, 2015.

[43] S. Deng, K.-Y. Liu, and X. Zhao. Semiparametric regression analysis of multivariate longitudinal data with informative observation times. *Comput Stat Data An*, 107:120–130, 2017.

[44] W. Su and H. Jiang. Semiparametric analysis of longitudinal data with informative observation times and censoring times. *J Appl Stat*, 45(11):1978–1993, 2018.

[45] J.F. Lawless and M. Zhan. Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *Can J Stat*, 26(4):549–565, 1998.

[46] H. Xie, G. McHugo, A. Sengupta, R. Clark, and R. Drake. A Method for Analyzing Longitudinal Outcomes with Many Zeros. *Mental Health Services Research*, 6(4):239–246, 2004.

[47] L. Liu, Y-C. T. Shih, R.L. Strawderman, D. Zhang, B.A. Johnson, and H. Chai. Statistical Analysis of Zero-Inflated Nonnegative Continuous Data: A Review. *Statistical Science*, 34(2):253–279, 2019.

[48] D. Lambert. Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics*, 34(1):1–14, 1992.

[49] L. Liu. Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Statistics in Medicine*, 28:972–986, 2009.

[50] V. Rondeau, E. Schaffner, F. Corbiere, J.R. Gonzalez, and S. Mathoulin-Pelissier. Cure frailty models for survival data: Application to recurrences for breast cancer and to hospital readmissions for colorectal cancer. *Statistical Methods in Medical Research*, 22(3):239–246, 2013.

[51] L. Liu, X. Huang, A. Yaroshinsky, and J.N. Cormier. Joint Frailty Models for Zero-Inflated Recurrent Events in the Presence of a Terminal Event. *Biometrics*, 72:204–214, 2016.

[52] Y-J. Kim. Joint model for bivariate zero-inflated recurrent event data with terminal events. *Journal of Applied Statistics*, 48(4):738–749, 2021.

[53] M-H. Chen, J.G. Ibrahim, and D. Sinha. A new joint model for longitudinal and survival data with a cure fraction. *Journal of Multivariate Analysis*, 91(1):18–34, 2004.

[54] L. Yang, H. Song, Y. Peng, and D. Tu. Joint analysis of longitudinal measurements and survival times with a cure fraction based on partly linear mixed and semiparametric cure models. *Pharmaceutical Statistics*, 20:362–374, 2021.

[55] T.M. Therneau, P.M. Grambsch, and T.R. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

[56] P.J.M. Verweij, H.C. van Houwelingen, and T. Stijnen. A Goodness-of-Fit Test for Cox's Proportional Hazards Model Based on Martingale Residuals. *Biometrics*, 54:1517–1526, 1998.

[57] O.O. Aalen, J. Fosen, H. Weedon-Fekjær, Ø. Borgan, and E. Husebye. Dynamic Analysis of Multivariate Failure Time Data. *Biometrics*, 60:764–773, 2004.

[58] P. Diggle, D. Farewell, and R. Henderson. Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. *Journal of Applied Statistics*, 56(5):499–550, 2007.

[59] E. Elgmati, D. Farewell, and R. Henderson. A martingale residual diagnostic for longitudinal and recurrent event data. *Lifetime Data Analysis*, 16:118–135, 2010.

[60] U.S. Govindarajulu, H. Lin, K.L. Lunetta, and R.B. D'Agostino Sr. Frailty models: Applications to biomedical and genetic studies. *Statistics in Medicine*, 30:2754–2764, 2011.

[61] J.M. Neuhaus and J.D. Kalbfleisch. Between- and Within-Cluster Covariate Effects in the Analysis of Clustered Data. *Biometrics*, 54:638–645, 1998.

[62] T. Lalonde. Modeling time-dependent covariates in longitudinal data analyses. In D.G. Chen and J. Wilson, editors, *Innovative Statistical Methods for Public Health Data.*, pages 57–79. Springer, Cham., 2015.

[63] T.L. Lai and D. Small. Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method-of-moments approach. *Journal of the Royal Statistical Society: Series B*, 69(1):79–99, 2007.

[64] T.L. Lalonde, J.R. Wilson, and J. Yin. Gmm logistic regression models for longitudinal data with time dependent covariates and extended classifications. *Statistics in Medicine*, 33:4756–4769, 2007.

[65] K. Kiernan, Jill Tao, and P. Gibbs. Tips and Strategies for Mixed Modeling with SAS/STAT ® Procedures. 2012.