1-1-2012

Statistical Methods for Non-Ignorable Missing Data With Applications to Quality-of-Life Data.

Kaijun Liao
University of Pennsylvania, kliao@mail.med.upenn.edu

Statistical Methods for Non-Ignorable Missing Data With Applications to Quality-of-Life Data.

Abstract
Researchers increasingly use more and more survey studies, and design medical studies to better understand the relationships of patients, physicians, their health care system utilization, and their decision making processes in disease prevention and management. Longitudinal data is widely used to capture trends occurring over time. Each subject is observed as time progresses, but a common problem is that repeated measurements are not fully observed due to missing response or loss to follow up. An individual can move in and out of the observed data set during a study, giving rise to a large class of distinct "non-monotone" missingness patterns. In such medical studies, sample sizes are often limited due to restrictions on disease type, study design and medical information availability. Small sample sizes with large proportions of missing information are problematic for researchers trying to understand the experience of the total population. The information in the data collected may produce biased estimators if, for example, the patients who don't respond have worse outcomes, or the patients who answered "unknown" are those without access to medical or non-medical information or care. Data modeled without considering this missing information may cause biased results.

A first-order Markov dependence structure is a natural data structure to model the tendency of changes. In my first project, we developed a Markov transition model using a full-likelihood based algorithm to provide robust estimation accounting for "non-ignorable" missingness information, and applied it to data from the Penn Center of Excellence in Cancer Communication Research. In my second project, we extended the method to a pseudo-likelihood based approach by considering only pairs of adjacent observations to significantly ease the computational complexities of the full-likelihood based method proposed in the first project. In my third project, we proposed a two stage pseudo hidden Markov model to analyze the association between quality of life measurements and cancer treatments from a randomized phase III trial (RTOG 9402) in brain cancer patients. By incorporating selection models and shared parameter models with a hidden Markov model, this approach provides targeted identification of treatment effects.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Epidemiology & Biostatistics

First Advisor
Andrea B. Troxel

Keywords
Clinical trial /Cancer applications, Composite likelihood /pseudo likelihood method, Conjoint analysis, Hidden Markov model, Longitudinal and multivariate data, Non-ignorable missing data

Subject Categories
Biostatistics

STATISTICAL METHODS FOR NON-IGNORABLE MISSING DATA WITH APPLICATIONS TO QUALITY-OF-LIFE DATA.

Kaijun Liao

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2012

Supervisor of Dissertation

Andrea B. Troxel, Professor of Biostatistics

Graduate Group Chairperson

Daniel F. Heitjan, Professor of Biostatistics

Dissertation Committee

Mary E. Putt, Associate Professor of Biostatistics

Benjamin C. French, Assistant Professor of Biostatistics

Katrina Armstrong, Professor of Medicine

STATISTICAL METHODS FOR NON-IGNORABLE MISSING DATA WITH APPLICATIONS TO QUALITY-OF-LIFE DATA.

ACKNOWLEDGEMENT

Completing a PhD is truly like running a marathon, and I would not have been able to complete this journey without the support of many people. I must first express my gratitude to my advisor, Prof. Andrea B. Troxel, as an outstanding advisor and excellent professor. She provided me many insightful comments and advice, constant encouragement, and invaluable suggestions which guided me through all the challenges and difficulties during the entire process of completing my doctoral work. What's most important is that she taught me how to be an excellent statistician and this experience will definitely benefit me for my entire career. I would like to thank all the members of my committee, Prof. Katrina Armstrong, Prof. Mary E. Putt, and Prof. Benjamin C. French, for their insightful comments, and their time and effort in reviewing and improving the quality of this work. I also feel lucky to have the opportunity to work with Prof. Katrina Armstrong's research team and collaborate on many levels of the research projects. I highly appreciate her knowledge, expertise, understanding, and patience, which added considerably to my graduate experience. I would like to thank Prof. Mary E. Putt for advising my master level project and all her encouragement, Prof. Nandita Mitra for her career suggestions and class selection during the graduate years, Prof. Benjamin C. French for reviewing my thesis and providing valuable feedback, and other faculty members at Penn who have offered me tremendous help throughout my Ph.D. years.

Finally, I am deeply and forever indebted to my parents and my wife for their love, support and encouragement throughout my entire life. Without their love and support, this dissertation would never have come into existence.

ABSTRACT

STATISTICAL METHODS FOR NON-IGNORABLE MISSING DATA WITH APPLICATIONS TO QUALITY-OF-LIFE DATA.

Kaijun Liao

Andrea B. Troxel

Researchers increasingly use more and more survey studies, and design medical studies to better understand the relationships of patients, physicians, their health care system utilization, and their decision making processes in disease prevention and management. Longitudinal data is widely used to capture trends occurring over time. Each subject is observed as time progresses, but a common problem is that repeated measurements are not fully observed due to missing response or loss to follow up. An individual can move in and out of the observed data set during a study, giving rise to a large class of distinct "non-monotone" missingness patterns. In such medical studies, sample sizes are often limited due to restrictions on disease type, study design and medical information availability. Small sample sizes with large proportions of missing information are problematic for researchers trying to understand the experience of the total population. The information in the data collected may produce biased estimators if, for example, the patients who don't respond have worse outcomes, or the patients who answered "unknown" are those without access to medical or non-medical information or care. Data modeled without considering this missing information may cause biased results.

A first-order Markov dependence structure is a natural data structure to model the tendency of changes. In my first project, we developed a Markov transition model using a full-likelihood based algorithm to provide robust estimation accounting for "non-ignorable" missingness information, and applied it to data from the Penn Center of Excellence in Cancer Communication Research. In my second project, we extended the method to a pseudo-

likelihood based approach by considering only pairs of adjacent observations to significantly ease the computational complexities of the full-likelihood based method proposed in the first project. In my third project, we proposed a two stage pseudo hidden Markov model to analyze the association between quality of life measurements and cancer treatments from a randomized phase III trial (RTOG 9402) in brain cancer patients. By incorporating selection models and shared parameter models with a hidden Markov model, this approach provides targeted identification of treatment effects.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : Introduction

In chronic disease studies, questionnaires are an often primary source of information to measure changes in attitude or compliance with treatment or medical advice. More and more survey studies focus on questionnaires of patients with different health issues, stages of disease, types of cancer, and other medical/non-medical information so that health providers or decision makers can better understand patient behavior and the estimates of treatment effects. The underlying structure of the quality and quantity of information that can be collected from each participant can be complicated due to the fact that during follow-up, the occurrence of observations at a given time depends on many observed or unobserved factors. Intuitively, patient behavior involves attitudes and knowledge. So questionnaires, health-related attitudes and information clearly are relevant. It is reasonable to expect that patients' responses could be lower for those with worse health, or could be a function of all health information, such as disease type, how actively patients seek medical help, and their supporting environment; this makes the missingness more likely to be informative. Longitudinal data is widely used to monitor disease progression, or investigate changes over time in a characteristic which is measured repeatedly for each study participant. Missing information is typically inevitable in longitudinal studies, and can result in biased estimates and a loss of power when the missingness is informative.

In Chapter 2, we propose a full-likelihood based transition model and apply it to data from the Penn Center of Excellence in Cancer Communication Research, a cancer-related survey study recently conducted at the University of Pennsylvania. One of the research goals of the study was to examine how the Patient-Clinician Information Engagement (PCIE) score affects cancer patients' attitudes and behaviors in breast, prostate, and colorectal cancers; in particular, researchers were interested in the amount of exercise the patients were engaged in. Decisions people choose to follow will impact their health status. For example, patients decide whether to increase exercise, to get radiation therapy, or to choose surgery after seeking out treatment information from their physicians. The decision making process may

be influenced by both medical and non-medical information. A random sample was selected in fall 2006 from the Pennsylvania Cancer Registry (PCR). Patients had to have one of the above three cancers, diagnosed in 2005. There were a total of 2010 cancer patients who responded to at least one of three surveys, including 650 patients with prostate cancer, 682 patients with colorectal cancer, and 678 patients with breast cancer. The study included three longitudinal surveys. Surveys were initially conducted in fall 2006, with the second and third waves conducted in fall 2007 and fall 2008. The response rate for PCIE scores were 99.00% for wave one, 63.28% for wave two, and 55.67% for wave three. Clearly this study resulted in a large amount of missing data for unknown reasons, and thus requires careful attention to the issue of missingness.

We use a full-likelihood based method to analyze continuous longitudinal responses with non-ignorable non-monotone missing data, and consider a transition probability model for the missingness mechanism. A first-order Markov dependence structure is assumed for both the missingness mechanism and observed data. This process fits the natural data structure in the longitudinal framework. Instead of using logistic regression to model the missing mechanism, we propose a beta-binomial distribution to model the probability of non-response. The beta-binomial distribution can be extended to the multivariate Polya distribution when there are more than two types of responses; our main interest is in estimating the parameters of the marginal model and evaluating the MAR (missing at random) assumption in the Effects of Public Information Study. We also present a simulation study to assess model performance in small samples, addressing the basic issues of bias in the parameter estimates and computing coverage probabilities, while varying the covariance structure of the longitudinal outcomes. The marginal effects are estimated well even when the underlying data distribution is not normal. However, full-likelihood based methods require integration over the unobserved data. The parameter estimation has to be done numerically, and this can be computationally prohibitive due to the complicated joint likelihood function, especially when the number of repeated assessments is large.

Pseudo-likelihood methods (Gong and Samaniego, 1981; Parke, 1986) and composite marginal likelihood methods (Cox and Reid, 2004; Varin et al., 2011) are widely used to ease the computational complexities of the conventional likelihood-based method. The pseudo-likelihood methods can be viewed as an extension of composite marginal likelihood methods, which can be transferred into the non-ignorable non-monotone missing data framework. In Chapter 3, we propose a pseudo-likelihood method based on the conditional density of all adjacent pairs of assessments, with a first-order auto-regressive covariance structure to account for the correlation of the repeated observations within subjects. Estimation proceeds using the pseudo-score vector, which guarantees a consistent estimator. Although the pseudo-likelihood method achieves asymptotically unbiased estimators of the regression parameters and missingness parameters if the model is correctly specified, these estimators can be highly inefficient in the case of faulty assumptions about the covariance structure across measurement times. A sandwich estimator is used to obtain correct inference for variance parameters. We fitted the proposed method to the same data from the Penn Center of Excellence in Cancer Communication Research as in project one. A simulation study investigates the empirical behavior of the proposed models, compared to the full-likelihood method proposed in Chapter 2. The simulation study shows that this approach can handle longitudinal data with various covariance structures well and is no more computationally intensive than the independent pseudo-likelihood model (Troxel et al., 1998b). This approach can handle a mis-specified correlation to some extent. In simulation studies with a variety of mis-specified correlation structures, the marginal effects and missingness effects consistently have high coverage probabilities as long as the correlation among pairs is nonzero.

In Chapter 4, we extend our approach using a hidden Markov model framework. By incorporating both selection models and shared parameter models, we can identify differences among the transition processes with incomplete data simultaneously in both a state-dependent model and a missingness mechanism model. The conditional independence assumed in the hidden Markov model provides a simple framework for reducing the

multi-dimensional integration in traditional methods into one dimensional integration in the observed likelihood. In addition, the proposed models avoid the problem of specification of the correlation structure of repeated outcomes by instead emphasizing estimation in Markov Chain parameters. We propose a generalized linear model and generalized linear mixed model framework, using a *Baum-Welch* algorithm (Baum et al., 1970; Rabiner, 1989; Welch, 2003) to update the Markov Chain parameters to provide efficient parameter estimation in the general situation of non-ignorable non-monotone longitudinal missing data. A two-stage pseudo-likelihood method is used to reduce the parameter space to make this model more attractive. Our proposed method is applied to data from a randomized phase III intergroup trial conducted by the Radiation Therapy Oncology Group (RTOG 9402) between 1994 and 2002, coordinated by the National Cancer Institute, in anaplastic oligodendroglioma (AO) brain tumor, patients received either chemotherapy plus radiation therapy (Arm 1) or radiation therapy alone (Arm 2), as previously described by Cairncross et al. (2006) and Wang et al. (2010). Previous reports had shown that AO patients respond to surgery and radiotherapy (RT) at diagnosis, as well as to procarbazine, lomustine, and vincristine (PCV) chemotherapy; it was unclear whether patients would benefit from combined PCV and RT therapy, compared to RT alone. Study reports also showed that patients who lack the 1p and 19q chromosomes have significantly longer progression survival times when treated with PCV+RT, but this is associated with substantial toxicity. In RTOG 9402, there was no significant difference in median survival times between the two treatment arms in patients with only one co-deletion or no deletions of chromosomes. The effect of toxicity and side effects from PCV chemotherapy and RT on patients' neurologic functioning and global quality of life remains unclear. Several measures were collected at each visit to assess patients cognitive ability and attitudes on quality of life during the study time period, including Karnofsky performance status (KPS), which measures physical well-being; the Mini-Mental Status Exam (MMSE), which measures cognitive ability as assessed by a nurse, research associate, or physician to reflect the opinions of the health care specialist; and the modified Brain Quality of Life Questionnaire (B-QLQ), which mea-

sures patient-reported quality of life. In this Chapter, we focus on the association between patients' MMSE/B-QLQ scores and treatment effect. By modeling the disease progression through different hidden states, our approach allows more precise identification of the treatment effects.

CHAPTER 2 : A transition model for quality of life data with non-ignorable non-monotone missing data

2.1. Introduction

In a longitudinal study, each subject is observed as time progresses. A common problem is that repeated measurements are not fully observed due to missing responses or loss to follow up. An individual can move in and out of the observed data set during the study, giving rise to a large class of distinct "non-monotone" missingness patterns. The appropriate statistical methods differ based on the nature of the data structure and missing mechanism. The simplest types of incomplete data are when the missingness is MCAR (missing completely at random) or MAR (missing at random). Little and Rubin (1987) and Allison (2001) provide helpful terminology to describe missing data mechanisms and a comprehensive overview of methods in this setting. Most approaches can be categorized as selection models, pattern-mixture models or shared-parameter models depending on the factorization of the joint likelihood of the outcomes and missingness indicators. This article will focus on selection models.

Under the MCAR mechanism, the observed data can be viewed as a random subset of the complete data. For the MAR assumption, the missingness mechanism depends only on observed quantities. Both mechanisms can be treated as "ignorable" if the parameters in the two parts of the model are distinct. For "ignorable" data, generalized estimating equations (GEE) provide asymptotic unbiased estimation if the underlying data is MCAR (Liang and Zeger, 1986). Weighted generalized estimating equations (WGEE) can provide unbiased estimation if the underlying data is MAR (Robins and Rotnitzky, 1995). However, none of above methods can provide consistent unbiased estimators under informative dropout or non-ignorable missingness. The approaches to modeling informative drop out or non-ignorable missing data in the longitudinal setting depend on the nature of the data structure, data type, variance/covariance structure, and proportion of missing data. Many

proposed methods assume a multivariate Gaussian distribution for the outcomes, with different specifications of the covariance structure; these include (Verbyla and Cullis, 1990; Richard and Lynn, 1990; Munoz et al., 1992; Diggle and Kenward, 1994). Diggle and Kenward (1994) proposed a likelihood-based method for continuous longitudinal outcomes with non-ignorable or informative drop-out. They specified a multivariate Gaussian distribution for the data and a logistic model for the probability of missing observations. Their model allowed the missingness probability to depend on previous and current measurements, and the likelihood was integrated over the range of the unobserved values. The likelihood involved approximations with numerical integration and iterative computations. However, their method required monotone missingness, also called informative drop-out.

Troxel et al. (1998a) extended the method to allow a non-monotone and non-ignorable missingness mechanism. They proposed a logistic model that allowed the probability of non-response to depend on the value of the current and/or previous measurement, allowing for a non-ignorable missing data mechanism, and assumed multivariate Gaussian distribution for the underlying outcomes. They assumed a first-order Markov dependence structure to facilitate estimation.

Another way to attack the problem of non-ignorable non-monotone missingness in longitudinal data is using pseudolikelihood methods to greatly ease the computational burdens of the full-likelihood method, by setting the nuisance parameter at zero or some convenient estimate. Troxel et al. (1998b), Sinha et al. (2010), and Parzen et al. (2007) used pseudolikelihood methods to deal with the binary case. Troxel et al. (2010) used an optimal weighted combination of two pseudolikelihoods to increase the efficiency of the estimation. Tsonaka et al. (2009) considered a semi-parametric shared parameter model without assuming any parametric assumption for the random effects distribution.

Our method is an extension of the work of Troxel et al. (1998a). As in the earlier work we adopt the multivariate Gaussian distribution assumption for the underlying data and the first-order Markov dependence structure. Instead of using a logistic regression to model

the missing mechanism, we propose a beta-binomial distribution to model the probability of non-response. The multivariate Polya distribution is a high-dimensional version of the beta-binomial distribution; the beta and binomial distributions correspond to Dirichlet and multinomial distributions, respectively, in the multivariate situation. Because of this property, our approach can be easily extended into more than one state of missingness, such as intermediate missingness, drop-out or even death if there is non-response due to death. Because of the Gamma function and/or Beta functions involved, closed-form maximum likelihood estimates are impractical. We propose to use Gauss-Hermite quadrature as suggested in Liu and Pierce (1994) to approximate the likelihood. The Broyden-Fletcher-Goldfarb-Shanno (BFGS) (Nocedal and Wright, 2006) algorithm is applied to search for optimal solutions. The beta-binomial model provides superior model fitting to the data compared to a traditional logistic model, especially for binary data with unbalanced sparse data. From a Bayesian perspective, the beta is the conjugate prior distribution for the parameters of the binomial distribution. The parameters α and β of the beta distribution can be thought of as pseudo-observations of "success" and "failure" to be added to the actual number of successes or failures observed. This helps to stabilize the estimation of the missingness mechanism, especially when some time points have small amounts of missing or no missing data. This mixture model also reduces multimodality in the likelihood.

The proposed methods were applied to the data from the Penn Center of Excellence in Cancer Communication Research. Effectiveness of communication between patients and their physicians is a very important factor in cancer research, and throughout the health care system. Effective exchange of information between patients, physicians, health care systems, and the environment surrounding them determines how active participants are within the health care system. There are many studies showing a link between highly isolated areas or individuals and worse outcomes in cancer research (Putt et al., 2009), including shorter survival time, worse quality of life, and lower rates of participation in recommended treatment programs. The rate of patient adherence to a recommended course of treatment is normally higher in patients who actively seek information about their cancer

treatment and quality of life from different channels (Tan et al., 2011). So it is crucial to understand the relationship between patients, their physicians, and the health care system around them, as well as the role of shared decision-making skills; how patients get, give, and discuss information and make health care decisions is important in cancer research, especially given the high demands that the healthcare system is facing.

There are a total of 2010 cancer patients who responded to at least one of three surveys, including 650 patients with prostate cancer, 682 patients with colorectal cancer, and 678 patients with breast cancer. The study included three longitudinal surveys. Surveys were initially conducted in fall 2006, with the second and third waves conducted in fall 2007 and fall 2008. The response rates of possible explanatory variables are listed in Table 2.2. Clearly this study resulted in a large amount of missing data for unknown reasons, may have an important impact on inference derived from this study.

The study sample was randomly selected in fall 2006 from the Pennsylvania Cancer Registry (PCR). Patients had to have one of the above three cancers, diagnosed in 2005. The American Association for Public Opinion Research (AAPOR, 2006) response rates for the primary sample were 68%, 64%, and 61% for the respective cancer groups (Nagler et al., 2010). Surveys were mailed to all participants using Dillman's design method (Dillman, 2010). All patients were first mailed an introductory letter explaining the purpose of the study and including instructions; the surveys were mailed in a subsequent packet with a small monetary incentive ($3 or $5 for the short or long version of the survey). Reminder letters were sent after 2 weeks for subjects who did not return the survey. Patient consent was provided prior to participation, and the University of Pennsylvania Institutional Review Board reviewed and approved this study.

One of the research goals of the study described here is to examine how the Patient-Clinician Information Engagement (PCIE) score affects breast, prostate, and colorectal cancer patients' attitudes and behaviors; in particular, researchers were interested in the amount of exercise the patients engaged in. Decisions people choose to follow will impact their health

status. For example patients decide whether to increase exercise, to get radiation therapy, or to choose surgery after seeking treatment information from their physicians. The decision making process may be influenced by both medical and non-medical information. PCIE scores are measured from 8 items; for each item, patients think back to the first few months of their cancer diagnosis and recall whether they have 1) sought information about treatments from their treating physician; 2) sought treatment information from other physicians or health professionals; 3) actively looked for information about their cancer from their treating physician; 4) actively looked for information about their cancer from other physicians or health professionals; 5) discussed information from other sources with their treating physician; 6) received suggestions from their treating physician to get information from other sources; 7) actively looked for information about quality of life issues from their treating physician; and 8) looked for quality of life information from other physicians or health professionals. Each of the eight items was transformed to a Z-score, and the average of the eight Z-scores formed the PCIE scale.

We use the extent of exercise ("During an average week, how many days do you exercise?") as the primary outcome. The outcomes range from 0 to 7 by experimental design; we treat these as continuous responses in this small interval. The Pearson correlation coefficients in Table 2.4 suggest that the correlation between baseline and follow-up is greater than the correlation among the follow-up assessments. We use the unstructured correlation in the data analysis and simulation sections, and we extend the correlation into AR(1), exchangeable and Toeplitz later in the simulation section for further model assessment.

The proposed methods are described in Section 2.2, and illustrated with an analysis of the PCIE data in Section 2.3. A simulation study to address the performance of the methods is presented in Section 2.4. Section 2.5 provides a discussion and ideas for future work.

2.2. Methods and Notation

2.2.1. Notation and underlying assumptions

Given a longitudinal data set, let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iT})'$ represent the vector of repeated measurements for subject i $(i = 1, \ldots, n)$ with T measurement times. Let $\mathbf{X}_i$ be a vector of p covariates observed on the ith subject. The covariate vector $\mathbf{X}_i$ could be either time independent or time dependent. Because the repeated measurements are not fully observed at each time point $t = (1, \ldots, T)$, define a vector of missingness indicators $\mathbf{R}_i = (R_{i1}, R_{i2}, \ldots, R_{iT})$ to correspond with the outcome vector $\mathbf{Y}_i = (\mathbf{Y}_{i,obs}, \mathbf{Y}_{i,mis})$. Each element of R_i is defined as

$$R_{it} = \begin{cases} 0 & \text{if missing} \\ 1 & \text{if observed} \end{cases}.$$

For each subject, the full data are given by the repeated measurements and missingness indicators with joint distribution $L(\theta, \beta | \mathbf{Y}_i, \mathbf{R}_i, \mathbf{X}_i) \propto P(\mathbf{Y}_i, \mathbf{R}_i | \mathbf{X}_i, \theta, \beta)$. By partitioning $\mathbf{Y}_i$ into $(\mathbf{Y}_{i,obs}, \mathbf{Y}_{i,mis})$, we can rewrite the joint likelihood in several ways. θ is parameter space associated with outcome process, and β is parameter space associated with missingness mechanism. A selection model would specify the joint distribution using the marginal distribution of the repeated outcomes and the conditional distribution of missing indicators:

$$P(\mathbf{Y}_i, \mathbf{R}_i | \mathbf{X}_i, \theta, \beta) = P(\mathbf{Y}_{i,obs}, \mathbf{Y}_{i,mis} | \mathbf{X}_i, \theta) P(\mathbf{R}_i | \mathbf{Y}_{i,obs}, \mathbf{Y}_{i,mis}, \mathbf{X}_i, \beta).$$

A pattern-mixture model assumes the full data have different distributions across strata determined by the pattern of missingness:

$$P(\mathbf{Y}_i, \mathbf{R}_i | \mathbf{X}_i, \theta, \beta) = P(\mathbf{R}_i | \mathbf{X}_i, \beta) P(\mathbf{Y}_{i,obs}, \mathbf{Y}_{i,mis} | \mathbf{R}_i, \mathbf{X}_i, \theta).$$

A shared-parameter model assumes independence between the complete data and missing

indicators conditional on group of shared parameters $\boldsymbol{\gamma}$:

$$P(\mathbf{Y}_i, \mathbf{R}_i | \mathbf{X}_i, \theta, \beta) = \int P(\mathbf{Y}_{i,obs}, \mathbf{Y}_{i,mis} | \gamma_i, \mathbf{X}_i, \theta) P(\mathbf{R}_i | \gamma_i, \mathbf{X}_i, \beta) p(\gamma_i) d\gamma_i.$$

In our study, we focus on selection models, which are a natural way to factor the joint likelihood function. The diagram below indicates the relationships among the variables graphically. Each line indicates the dependence between the nodes.

$$\begin{array}{ccccccc} Y_{i1} & \longrightarrow & Y_{i2} & \cdots & Y_{i,T-1} & \longrightarrow & Y_{iT} \\ \downarrow & \cdots & \downarrow & \cdots & \downarrow & \cdots & \downarrow \\ R_{i1} & \longrightarrow & R_{i2} & \cdots & R_{i,T-1} & \longrightarrow & R_{iT} \end{array}$$

We adopt a similar model to Troxel et al. (1998a), and assume $\mathbf{Y}_i \sim MVN(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where the mean structure $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \cdots \mu_{iT})$ depends on a p-dimensional covariate vector $\mathbf{X}_i$. We also assume a first-order Markov dependence structure for both the full outcome data and the missingness indicators, so that $f(Y_{it}|Y_{i1}, Y_{i2}, \ldots Y_{it-1}) = f(Y_{it}|Y_{it-1})$ and $f(R_{it}|R_{i1}, R_{i2}, \ldots R_{it-1}) = f(R_{it}|R_{it-1})$. Let $\sigma_t^2 = var(Y_{it})$ and $\rho_t = corr(Y_{it}, Y_{it+1})$. Then we can denote the conditional likelihood as

$$Y_{it}|Y_{i,t-1} \sim N\Big\{\mu_{it} + \rho_{t-1}\frac{\sigma_t}{\sigma_{t-1}}(Y_{i,t-1} - \mu_{i,t-1}), \sigma_t^2(1-\rho_{t-1}^2)\Big\}.$$

For $T = 3$ the first-order ante-dependence structure is denoted as :

$$\Sigma = \left\{ \begin{array}{ccc} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_1\rho_2 \\ \sigma_2\sigma_1\rho_1 & \sigma_2^2 & \sigma_2\sigma_3\rho_1 \\ \sigma_3\sigma_1\rho_1\rho_2 & \sigma_3\sigma_2\rho_2 & \sigma_3^2 \end{array} \right\}.$$

2.2.2. Missingness mechanism model

Unlike other approaches to modeling the missingness mechanism, we are interested in the transition probability of the missingness indicators R_{it}. Conditional on each time t, the missingness mechanism becomes a two-state Markov chain. We model the transition probabilities $\pi_{jk} = Pr(R_{it} = j|R_{i,t-1} = k, Y_{it}, X_{it})$, $j = 0, 1; k = 0, 1$ as

$$\begin{pmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{pmatrix}$$

which satisfy the equation $\pi_{00} + \pi_{01} = \pi_{10} + \pi_{11} = 1$. We assume that the initial state is independent, and define $n_{j,k}$ as the number of times in the whole sequence that k is followed by j:

$$n_{j,k} = \sum_{t=1}^{T} I(R_t = j|R_{t-1} = k)$$
$$n_{j.} = \sum_k n_{j,k}, \quad n_{.k} = \sum_j n_{j,k}.$$

Then the missingness mechanism can be written as

$$\begin{aligned} \mathbb{L}_i &= \pi_{00}^{n_{i00}} \pi_{01}^{n_{i01}} \pi_{10}^{n_{i10}} \pi_{11}^{n_{i11}} \\ &= \prod_{t=2}^{T} \prod_{j=0}^{1} \prod_{k=0}^{1} \pi_{jk}(t)^{I(R_{i,t}=j|R_{i,t-1}=k)}. \end{aligned}$$

This becomes a product of binomial distributions. Logistic regression has been used for this type of problem but yield unstable estimates for binary outcomes near the boundary of the parameter space. Thus we estimate the probability of missingness at each time t using a joint beta-binomial distribution.

Given time $t-1$, the missingness mechanism follows $(R_{i,t}|R_{i,t-1} = k) \sim Bernoulli(\pi_{ikt})$; we impose a beta distribution on the missingness probability, $\pi_{ikt} \sim Beta(a_{ikt}, b_{ikt})$ Then we

have

$$f(R_{i,t}|R_{i,t-1}, y_{it}, \pi) = \prod_{k=0}^{1} \pi_{k1}^{I(R_{it}=1)I(R_{it-1}=k)}(1-\pi_{k1})^{[1-I(R_{it}=1)]I(R_{it-1}=k)}$$

$$f(\pi_{k1}|a_{k1}, b_{k1}) = \frac{\Gamma(a_{k1}+b_{k1})}{\Gamma(a_{k1})\Gamma(b_{k1})} \times \pi_{k1}^{a_{k1}-1}(1-\pi_{k1})^{b_{k1}-1}.$$

Integrating the π out, the mixture function can be expressed as

$$\begin{aligned} & f(R_{i,t}|R_{i,t-1}, a_{ik1}, b_{ik1}, y_{it}) = \int_0^1 f(R_{i,t}|R_{i,t-1}=k, \pi_{ikt}) f(\pi_{ikt}|a_{ikt}, b_{ikt}, y_{it}) d\pi_{ikt} \\ = & \prod_{k=0}^{1} \frac{\Gamma(a_{ik1}+b_{ik1})}{\Gamma(a_{ik1})\Gamma(b_{ik1})} \\ \times & \frac{\Gamma(a_{ik1}+I(R_{i,t}=1)I(R_{i,t-1}=k))\Gamma(b_{ik1}+[1-I(R_{i,t}=1)]I(R_{i,t-1}=k))}{\Gamma(a_{ik1}+b_{ik1}+I(R_{it-1}=k))} \end{aligned}$$

with $a_{ik1} = exp(\zeta_1 \mathbf{X_{it}} + \vartheta_\mathbf{1} \mathbf{Y_{it}} + \psi_\mathbf{1} \mathbf{R_{i,t-1}})$ and $b_{ik1} = exp(\zeta_2 \mathbf{X_{it}} + \vartheta_\mathbf{2} \mathbf{Y_{it}} + \psi_\mathbf{2} \mathbf{R_{i,t-1}})$. However, the link function chosen could be different resulting in a different missingness mechanism model. For given $R_{i,t-1} = 0$, the transition probability can be denoted as

$$P(R_{it} = l|R_{it-1} = 0, Y_{it}, X_{it}) : \quad = \begin{cases} \frac{1}{1+\exp((\zeta_1-\zeta_2')\mathbf{X_{it}}+(\vartheta_\mathbf{1}-\vartheta_\mathbf{2})\mathbf{Y_{it}})} & \text{if} \quad l=1 \\ \frac{1}{1+\exp(-(\zeta_1-\zeta_2')\mathbf{X_{it}}-(\vartheta_\mathbf{1}-\vartheta_\mathbf{2})\mathbf{Y_{it}})} & \text{if} \quad l=0 \end{cases}.$$

For given $R_{i,t-1} = 1$,

$$P(R_{it} = l|R_{it-1} = 1, Y_{it}, X_{it}) : \quad = \begin{cases} \frac{1}{1+\exp((\zeta_1-\zeta_2)\mathbf{X_{it}}+(\vartheta_\mathbf{1}-\vartheta_\mathbf{2})\mathbf{Y_{it}}+(\psi_\mathbf{1}-\psi_\mathbf{2}))} & \text{if} \quad l=1 \\ \frac{1}{1+\exp(-(\zeta_1-\zeta_2)\mathbf{X_{it}}-(\vartheta_\mathbf{1}-\vartheta_\mathbf{2})\mathbf{Y_{it}}-(\psi_\mathbf{1}-\psi_\mathbf{2}))} & \text{if} \quad l=0 \end{cases}.$$

Notice that if $\vartheta_1 - \vartheta_2 \neq 0$ then the missingness mechanism is indeed non-ignorable since the probability of missingness depends on the unobserved outcome Y_{it}. In practice, only the difference of each parameters are identifiable, not the individual parameters. We let $\zeta_c = \zeta_1 - \zeta_2$, $\vartheta_c = \vartheta_1 - \vartheta_2$ and $\psi_c = \psi_1 - \psi_2$ be the final parameters in the missingness

mechanism model, where ζ_c is the coefficient of the covariates, ϑ_c is the coefficient of the current observation y_{it} ,and ψ_c is the coefficient of the previous missingness indicator r_{it-1}.

The link function for parameters a_{ikt} and b_{ikt} of the Beta distribution could be chosen differently than a simple exponential function, and this will result a different missingness mechanism model. The missingness mechanism model could be expanded, similar to Dirichlet-Multinomial distribution, from the current Beta-Binomial distributions when modeling a missing data indicator with more than two levels, such as "observed", "intermittently missing", "drop out".

2.2.3. Parameter estimation

The observed joint likelihood function can be denoted as

$$\begin{aligned}
\mathcal{L}_i(\mu, \Sigma, \theta, \beta) &= f(\mathbf{Y}_{i,obs}, \mathbf{R}_i) \\
&= \int \dots \int f(\mathbf{Y}_{i,obs}, \mathbf{Y}_{i,miss}, \mathbf{R}_i) d\mathbf{Y}_{i,miss} \\
&= \int \dots \int f(Y_{i1}) f(R_{i1}|Y_{i1}) \prod_{t=2}^{T} f(Y_{it}|Y_{i,t-1}) f(R_{it}|R_{i,t-1}, Y_{it}) \mathbf{d}\mathbf{Y}_{i,miss}.
\end{aligned}$$

There is no closed form for the observed likelihood function due to the complicated joint likelihood; a numerical integration method will be applied to approximate the likelihood function. The Gauss-Hermite quadrature rule is defined as

$$\begin{aligned}
\int_{\mathbb{R}} f(t) d\lambda(t) &= \int_{\mathbb{R}} f(t) w(t) dt \\
&= \int_{\mathbb{R}} f(t) \exp(-t^2) dt = \sum_{k=1}^{m} w_k f(\tau_k) + R_m(f)
\end{aligned}$$

where m is the number of nodes, $d\lambda(t) = w(t)dt = exp(-t^2)dt$ is the measure with bounded or unbounded support on $\mathbb{R}$, w_k is the weight of the Gauss-Hermite quadrature rule, τ_k are the nodes (zero roots of the m^{th} order Hermite polynomials) and $R_m(f)$ is the error term. The τ_k are symmetric about zero. The error term $R_m(f)$ will be zero if $f(t)$ is

polynomial with degree less than $2m-1$. Let $\phi(t;\mu,\sigma)$ be a normal density with mean μ and standard deviation σ. Then for any given function $f(t)$ we can approximate an integral as a summation following the transfomation used in Liu and Pierce (1994).

$$\int_{\infty}^{\infty} f(t)\phi(t;\mu,\sigma)dt \simeq \sum_{i=1}^{m} \frac{w_i}{\sqrt{\pi}} f(\mu+\sqrt{2}\sigma\tau_k).$$

For $T=3$, we list all possible data patterns and the joint likelihood function in the Appendix.

Our model is likelihood based, so maximum likelihood theory holds for parameter estimation. Letting $\eta=(\mu_1,\mu_2,\cdots,\mu_T,\sigma_1,\sigma_2,\cdots,\sigma_T,\rho_1\rho_2,\cdots,\rho_{t-1},,\beta,\theta,\psi)$, we have

$$\sqrt{n}(\hat{\eta}-\eta_0) \sim \mathbf{MVN}(\varnothing, \mathcal{I}^{-\mathbf{1}})$$

The Fisher information matrix $\mathcal{I}$ is estimated using the observed information matrix $\hat{\mathcal{I}}$. The Hessian matrix can be calculated during the maximization step, and the inverted Hessian matrix provides the observed Fisher information matrix .

2.3. Example: Analysis of PCIE Data

More and more survey studies focus on questionnaires returned by patients with different health issues, stages of disease, type of cancer, and other medical/non-medical characteristics, so that health providers and/or decision makers can better understand the changing behavior of the patients. Intuitively, patients' behaviors involving attitude change and information seeking, as well as their propensity to respond to questionnaires, can be health-related. It is reasonable to expect that patients are less likely to respond in cases of worsened health, or that response propensity is a function of all health information, such as disease type, how actively subjects seek medical help, and how they are affected by their supporting environment, which makes the missingness more likely informative.

Table 2.1 lists the 8 missingness patterns in the PCIE data. In practice, pattern 1, in which

subjects are missing data at all three waves, carries no information and will be excluded from the study. We use the extent of exercise ("During an average week, how many days do you exercise?") as the primary outcome. The outcomes range from 0 to 7 by experimental design; we treat these as continuous responses in this small interval. There were 85.66% of patients who responded to the baseline survey, 61.75% who returned the survey in wave 2, and 56.03% who answered the questions in wave 3. We calculate the Pearson correlation coefficients, shown in Table 2.4, which suggests that the correlation between the baseline and follow-up assessments is greater than the correlation among the follow-up assessments. We use the unstructured correlation for data analysis and in the simulation section, and we extend the correlation into AR(1), exchangeable and Toeplitz later in the simulation section for further model assessment.

Table 2.3 lists all patient characteristics of interest for both the marginal model and the missingness model. There are a total of 2010 cancer patients who responded to at least one of three surveys, including 650 patients with prostate cancer, 682 patients with colorectal cancer, and 678 patients with breast cancer. The study included three longitudinal surveys. The cohort includes both male and female whose cancer stage ranges from mild (stage 0) to severe (stage 4). The age at cancer diagnosis ranges from a minimum of 23 to a maximum of 103. PCIE scores are measured from 8 items; for each item, patients think back to the first few months of their cancer diagnosis and recall whether they have 1) sought information about treatments from their treating physician; 2) sought treatment information from other physicians or health professionals; 3) actively looked for information about their cancer from their treating physician; 4) actively looked for information about their cancer from other physicians or health professionals; 5) discussed information from other sources with their treating physician; 6) received suggestions from their treating physician to get information from other sources; 7) actively looked for information about quality of life issues from their treating physician; and 8) looked for quality of life information from other physicians or health professionals. Each of the 8 items was transformed to a Z-score, and the average of the 8 Z-scores formed the PCIE scale. The summary table provides the variation of the

PCIE score at each times.

The parameters are estimated using the proposed method and compared to a GEE model, which assumes MCAR missingness, and a WGEE (weighted GEE), which assumes MAR. Both GEE and WGEE assume an "ignorable" mechanism. The traditional PCIE study considered the missingness mechanism as either MAR or MCAR, possibly resulting in biased results. We modify the standard GEE to address missingness in the data. The weighting is calculated using the missingness mechanism model first, followed by inversion of the observed probability to form the corresponding weights. The missingness mechanism model used "cancer type", "gender", "age at diagnosis", "cancer severity", "PCIE score" and the previous missingness indicator to predict the current missingness indicator. For "ignorable" data, WGEE will have unbiased estimators if the underlying data is MCAR. GEE will have a biased estimators if the underlying data is MAR.

Because missing covariate data was not of primary interest, a multiple imputation method was used to complete the missing covariates. Rubin (1987) proposed a multiple imputation method using a Monte Carlo approach in which the missing values are replaced by $m > 1$ simulated versions. We generated $m = 20$ replicates in our study. Each of the imputed datasets is analyzed using the proposed method, the GEE model and two weighted GEE models. The combined parameter estimates and confidence intervals from the $m = 20$ data sets follows Rubin's (Rubin, 1987) multiple imputation rule.

In Table 2.5, we list the parameter estimates after combined 20-fold imputation. The coefficient for Y_i in the missingness model indicates if the probability of missingness is related to the potentially unobserved values of the outcomes. A significant effect indicates that the longitudinal data is "non-ignorable". The coefficient for R_{i-1} indicates if the previous response had an effect on patients current response. $R_{i-1} = 1$ means that the previous response was collected. Clearly there are statistically significant effects in the missingness model for the coefficient of both Y_i $[-0.136(-0.216, -0.055)]$ and R_{i-1} $[-0.794(-0.955, -0.633)]$, which indicates that the MCAR assumption is invalid. The coefficients for both Y_i and R_{i-1} are

negative, indicating an inverse relationship with the missingness indicator. Patients who exercise more tend to be more likely to respond to the survey. They also tend to answer the questionnaire if they already answered the previous one. Patients who have prostate cancer $[-0.302(-0.595, -0.008)]$ are more likely to return the questionnaire. Severe cancer stage (stage 4) $[0.713(0.196, 1.230)]$ increases the missingness rate, which indicate that patients with advanced disease are less likely to respond to the survey. "Wave" has coefficient $[0.287(0.138, 0.424)]$ which suggests that patients tend to be less responsive to the survey as time increases; this happens typically in repeated measures studies; in that participants become less compliant as the study advances.

The marginal estimates from our proposed model are somewhat larger than the ones from either GEE or WGEE model. However, the significance levels are consistent between the models. Only "age at diagnosis" and "cancer stage" are statistically significant. "Age at diagnosis" has coefficient 0.010 $(0.003, 0.018)$ indicating that older patients engage in more exercise then younger patients. The coefficient of "cancer stage" $[-0.567(-1.075, -0.058)]$ indicates a negative correlation with outcome. Patients tend to reduce the amount of exercise when their cancer becomes more severe. PCIE did not show a statistically significant effect in either model which suggests we did not have enough evidence to show the patients's exercise behavior will be affected by differences in the PCIE score.

Although the MCAR and MAR assumption is apparently invalid, both GEE and WGEE models show similar trends to the proposed model; while most of the parameters estimates are attenuated, the inferential conclusions are unchanged in this example. The weighted GEE model provides similar results to the GEE model when the sample size large.

2.4. Simulation Study

2.4.1. Simulation results

In this section we use a simulation study to assess model performance in small samples, addressing the basic issues of bias in the parameter estimates and computing coverage

probabilities. We simulated N subjects with three potential measurement times, for $N = 300$, $N = 400$ and $N = 500$. For each setting of N, we increase the missing rate from low to high. In the low missingness situation, there is about 10% missing at time 1, $20\% - 25\%$ at time 2 , and around $40\% - 45\%$ missing at time 3. In the higher missingness setting, there is $30\% - 35\%$ missing at time 2 and $55\% - 60\%$ missing at time 3. The true parameters were selected to fit the proportion of each missingness pattern. Both the proposed method, the GEE model and weighted GEE model are applied in these six data settings. 500 simulations have been run to assess the model's performance; results are displayed in Table 2.6. The data were generated as trivariate normal, with pairwise correlation parameters $\rho_{1,2} = 0.4$ and $\rho_{2,3} = 0.2$. The variance for time 1 is $\sigma_1 = 1.2$, for time 2 is $\sigma_2 = 2.6$ and for time 3 is $\sigma_3 = 3.0$. The estimators are good for both the marginal parameter and missingness mechanism model in Table 2.6. The bias is very small. Both GEE and WGEE model consistently underestimate the parameters when the sample size is small, and the bias is substantial. This becomes much more severe when the missingness rate increases from low to high. For WGEE model, the weights are calculated through the missingness mechanism model first, and inverse of the observed probability forms the weights. "Intercept", "time" and previous missingness indicator R_{i-1} are used to predict the observed probability. Consistently, WGEE provides better estimators than GEE model across all data setting, although the bias are still substantial compared with the proposed model. WGEE model performs better when the missing proportion increases than does the GEE model.

2.4.2. Model Comparison

The proposed model is compared with the original model in Troxel et al. (1998a), which used the same settings for the complete data and a different logistic model for the missingness indicators denoted as $logit(\pi_{r_{it}=1}) = \beta_{0t} + \beta_1 Y_{it}$. In the Troxel et al. (1998a) model, this missingness model did not include the previous missing indicator as a covariate. We generated two data settings, one with our proposed model and one with the correctly spec-

ified Troxel et al. (1998a) model. The correctly specified original model from Troxel et al. (1998a) will become a misspecified model if the coefficient of the previous missing indicator is not zero. Our proposed model will be over-specified if the parameter of the previous missing indicator is zero. Table 2.7 shows these comparison results. When the parameter (ψ) of the previous missing indicator is not zero, the estimates from our transition model are unbiased and have high coverage probabilities. The Troxel 1998 model has good estimation in the marginal model and variance-covariance structure, but poor estimation in the missingness model. This is not surprising, since the missingness model is misspecified. When the parameter (ψ) of the previous missing indicator is zero, both models have very good estimation. The proposed model uses a small value to estimate the ψ with 95% coverage rate including zero.

Next, we fit the proposed model with three different covariance structures to see how our model handles a miss-specified correlation matrix. Our transition model uses ANTE(1) (ante-dependence) structure denoted as $\sigma_i\sigma_j \prod_{k=i}^{j-1} \rho_k$ for the $(i,j)th$ element. There are a total of $2t-1$ parameters needed. This will become computationally burdensome when t, the number of repeated times, increases. In practice the AR(1) (autoregressive(1)) structure is widely used, denoted as $\sigma^2\rho^{i-j}$ for the $(i,j)th$ element. There are only two parameters needed. The Pearson coefficient matrix in Table 2.4 shows the correlation between baseline and followup is 0.644, and 0.618 for followup2 and followup3. We also have the coefficients from Table 2.4 for $\rho 1,2 = 0.643(0.607, 0.678)$ and $\rho 1,2 = 0.624(0.586, 0.662)$ are statistics same which make the AR(1) structure reasonable choice. Another two correlation structures used for comparison are exchangeable ($\sigma^2[\rho 1(i \neq j) + 1(i = j)]$) structure and TOEP(2) (Banded Toeplitz $\sigma^2_{|i-j|+1} 1(|i-j| < 2)$) structure.

The AR(1), exchangeable, and TOEP(2) structure for $T = 3$ are written respectively as:

$$\Sigma = \left\{ \begin{array}{ccc} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 \end{array} \right\}_{AR(1)} ; \Sigma = \left\{ \begin{array}{ccc} \sigma^2 & \sigma^2\rho & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2\rho & \sigma^2 \end{array} \right\}_{Exch} ; \Sigma = \left\{ \begin{array}{ccc} \sigma^2 & \sigma_1 & 0 \\ \sigma_1 & \sigma^2 & \sigma_1 \\ 0 & \sigma_1 & \sigma^2 \end{array} \right\}_{TOEP(2)}$$

The comparison table is listed in Table 2.8. The proposed model can handle the AR(1) structure well since it is a special case of ANTE(1) structure. Our model still performs quite well in estimating the marginal effects and missingness coefficients for both exchangeable and TOEP(2) structure. The variances are estimated with high coverage probabilities. Both correlation estimates are less efficient than for the AR(1) model.

2.4.3. Non-Normal Data

In this section we compared the proposed model to each other with different underlying assumptions about the data distribution. We simulated two data sets with same true parameters with different distributions. One data set was simulated from a trivariate normal distribution. Another was simulated from a trivariate Gamma distribution. A Clayton copula, which is an asymmetric Archimedean copula, was used to generate the trivariate Gamma data. This dependence structure of trivariate Gamma followed an exchangeable correlation. We used Kendall's formula (Kendall, 1976) to assure the same covariance structure between trivariate normal and trivariate Gamma data. We generate three correlation structures with high ($\rho = 0.707$), low ($\rho = 0.5$) and zero ($\rho = 0$) intra-subject correlation with sample sizes $n = 300$ and $n = 500$ to examine the models' performance in Table 2.9. Our proposed model performed quite well when the data are normally distributed. The estimator becomes less efficient when the data are independent ($\rho = 0$), which makes sense since the missingness model is miss-specified and thus the model is over-fitted. The marginal effect and missingness models are still estimated well when the underlying data distribution is not normal. However, the correlations are poorly estimated, although we still have quite good estimation of the variance parameter. The estimation improves when the correlation

is strong and worsens when the correlation is weak in the dependent data.

2.5. Discussion

We have presented an extension of the full likelihood-based algorithm to handle non-monotone and non-ignorable missing data. We assume a first-order Markov structure in both the complete data and missingness mechanism which is a natural way to capture the correlation among repeated measurements in a longitudinal data framework. The estimation of marginal effects is generally robust to correct specification of the covariance matrix and missingness mechanism.

As with any model-based approach to non-ignorable missing data, the current approach is subject to unavoidable assumptions about the complete data distribution and the missing data mechanism. It is important to consider all substantive information about the area of application, prior experience with missing data in similar situations, and expert opinion about the mechanism of missing data when building such models. In many areas, enough knowledge and experience exists to justify the necessary assumptions, and the benefit in terms of bias reduction can be significant.

Our transition model can be easily extended to model more than two states such as dropout or intermittent missingness. The numerical integration provides an accurate approximation but at the cost of increased computational complexity. We sometimes encountered a multimodal likelihood surface in our study. A method to handle such surfaces is to choose a vector of starting values and use GEE estimates to get the starting point as close to the true values as possible. There are many classes of correlation structure; while we can not explore all of them, the proposed model can handle the situation of a mis-specified correlation as demonstrated by our simulations. The marginal effects and missingness effects are consistently estimated with high coverage probability as long as the intra-subject correlation is incorporated. For studies with more than 3 assessment times, it will be difficult to examine complex correlation structures due to the increase in the number of parameters in

the model.

There are increasing trends of more and more survey studies to better understand the relationship of patients, physicians, and the health care system. In most such studies, however, sample sizes are limited due to restrictions on cancer type, study design and medical information availability. Small sample sizes with large proportions of missing information become more and more concerning for researchers, and limit generalizability. When information is masked due to reasons relating to the patient-physician relationship, lower response rates in patients with worse outcomes due to the disease or to accessibility to medical information and care, special approaches are needed. If data are modeled without considering this informative missing data, seriously biased inference may result.

Table 2.1: Missingness patterns in PCIE study

Pattern	Number of case	Pattern	Number of case
0 0 0	166	1 0 0	457
0 0 1	26	1 0 1	118
0 1 0	77	1 1 0	231
0 1 1	221	1 1 1	714

Table 2.2: Response rate for possible outcome

Response Rate	Exercise	PCIE	Seeking
wave 1	75.62%	99.00%	98.76%
wave 2	61.84%	63.28%	63.63%
wave 3	53.68%	55.67%	55.87%

Table 2.3: Characteristics of covariates

Type of cancer	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Colorectal	682	33.93%	682	33.93%
Breast	678	33.73%	1360	67.66%
Prostate	650	32.34%	2010	100%

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Male	987	49.10%	987	49.10%
Female	1023	50.90%	2010	100%

Stage	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	129	6.42%	129	6.42%
0	182	9.05%	311	15.47%
1	355	17.66%	666	33.13%
2	798	39.70%	1464	72.84%
3	243	12.09%	1707	84.93%
4	303	15.07%	2010	100%

Age at Diagnosis	Mean	Median	Min	Max
	64.74	65	23	103

PCIE score at time	Mean	Median	Min	Max
Wave 1	0.00245	0.00811	-1.27416	1.24094
Wave 2	0.00063	-0.20160	-0.70182	1.88602
Wave 3	0.00167	-0.22578	-0.60372	2.04041

Table 2.4: Pearson correlation matrix of exercise score

Pearson Correlation Coefficients
Prob H0: ρ=0
Number of Observations

	Wave 1	Wave 2	Wave 3
Wave 1	1	0.6438	0.5848
		< .0001	< .0001
	1520	945	832
Wave 2	0.6438	1	0.6184
	< .0001		< .0001
	945	1243	935
Wave 3	0.5848	0.6184	1
	< .0001	< .0001	
	832	935	1079

Table 2.5: Longitudinal analysis of PCIE data

	Non-Ignorable missing model					Weighted GEE					GEE				
Variable	Estimate	S.E	95%Lower	95%upper		Estimate	S.E	95%Lower	95%upper		Estimate	S.E	95%Lower	95%upper	
Intercept	3.705	0.402	2.916	4.494	*	3.540	0.424	2.708	4.372	*	3.580	0.410	2.775	4.385	*
TypeCancer.Breast	0.012	0.141	-0.265	0.289		0.044	0.143	-0.237	0.325		0.006	0.139	-0.266	0.278	
TypeCancer.Prostate	0.189	0.155	-0.114	0.493		0.190	0.165	-0.135	0.514		0.171	0.160	-0.142	0.484	
Gender	-0.283	0.154	-0.585	0.019		-0.284	0.164	-0.606	0.038		-0.269	0.159	-0.581	0.042	
Dianosis Age	0.010	0.004	0.003	0.018	*	0.011	0.004	0.003	0.019	*	0.010	0.004	0.003	0.018	*
Stage.1	-0.108	0.211	-0.522	0.306		-0.117	0.218	-0.544	0.310		-0.103	0.213	-0.521	0.315	
Stage.2	0.031	0.218	-0.396	0.458		-0.006	0.230	-0.457	0.446		0.029	0.223	-0.408	0.465	
Stage.3	-0.360	0.246	-0.842	0.122		-0.357	0.248	-0.844	0.130		-0.346	0.245	-0.826	0.134	
Stage.4	-0.567	0.259	-1.075	-0.058	*	-0.485	0.275	-1.024	0.053	*	-0.525	0.259	-1.033	-0.018	*
Wave	-0.061	0.046	-0.151	0.029		0.041	0.030	-0.018	0.100		0.040	0.030	-0.018	0.098	
PCIE	-0.004	0.059	-0.120	0.112		-0.014	0.056	-0.124	0.096		-0.007	0.054	-0.112	0.098	
$\sigma(y_1)$	2.115	0.038	2.041	2.190											
$\sigma(y_2)$	2.116	0.043	2.032	2.199											
$\sigma(y_3)$	2.069	0.047	1.977	2.160											
$\rho 1, 2$	0.643	0.282	0.090	1.195											
$\rho 2, 3$	0.624	0.273	0.090	1.158											
Missing data model															
y_i	-0.136	0.041	-0.216	-0.056	*										
r_{i-1}	-0.794	0.082	-0.955	-0.633	*										
π_1	0.824	0.287	0.262	1.387											
Intercept	-0.491	0.419	-1.311	0.329											
TypeCancer.Breast	-0.170	0.121	-0.407	0.067											
TypeCancer.Prostate	-0.302	0.150	-0.595	-0.008	*										
Gender	-0.175	0.126	-0.422	0.073											
Diagnosis Age	0.011	0.004	0.004	0.019	*										
Stage.1	-0.208	0.249	-0.696	0.281											
Stage.2	-0.163	0.220	-0.593	0.268											
Stage.3	-0.046	0.250	-0.536	0.444											
Stage.4	0.713	0.264	0.196	1.230	*										
Wave	0.281	0.073	0.138	0.424	*										
PCIE	-0.027	0.260	-0.536	0.482											

Table 2.6: Simulation study 500 replicates

								N=300								
Response Rate	Time 1	Time 2	Time 3						Time 1	Time 2	Time 3					
	0.909	0.840	0.605						0.911	0.681	0.434					
		Full Liklihood			GEE		WGEE			Full Liklihood			GEE		WGEE	
	TRUE	E.est	C.P	Bias	E.est	Bias	E.est	Bias	TRUE	E.est	C.P	Bias	E.est	Bias	E.est	Bias
Intercept	6.800	6.793	0.950	0.007	6.308	0.492	6.393	0.407	5.600	5.604	0.948	0.004	4.996	0.604	5.134	0.466
Time	1.050	1.062	0.956	0.012	1.518	0.468	1.445	0.395	0.300	0.301	0.942	0.001	0.914	0.614	0.782	0.482
Missingness model																
Intercept	-0.800	-0.990	0.958	0.190					-0.800	-0.894	0.952	0.094				
Time	1.250	1.103	0.952	0.147					1.250	1.213	0.956	0.037				
$y_i(\vartheta_c)$	-0.400	-0.404	0.936	0.004					-0.400	-0.412	0.950	0.012				
$r_{i-1}(\psi_c)$	2.500	2.676	0.976	0.176					2.500	2.628	0.962	0.128				
π	2.314	2.326	0.934	0.012					2.314	2.363	0.946	0.049				
Correlation structure																
σ_1	0.182	0.181	0.950	0.002					0.182	0.179	0.952	0.004				
σ_2	0.956	0.953	0.944	0.003					0.956	0.953	0.948	0.002				
σ_3	1.099	1.095	0.952	0.004					1.099	1.099	0.946	0.001				
$\rho_{1,2}$	0.847	0.849	0.922	0.002					0.847	0.853	0.952	0.006				
$\rho_{2,3}$	0.405	0.414	0.940	0.009					0.405	0.408	0.938	0.002				
								N=400								
Response Rate	Time 1	Time 2	Time 3						Time 1	Time 2	Time 3					
	0.911	0.826	0.593						0.910	0.668	0.425					
		Full Liklihood			GEE		WGEE			Full Liklihood			GEE		WGEE	
	TRUE	E.est	C.P	Bias	E.est	Bias	E.est	Bias	TRUE	E.est	C.P	Bias	E.est	Bias	E.est	Bias
Intercept	6.400	6.402	0.942	0.002	5.913	0.487	6.001	0.399	5.400	5.402	0.930	0.002	4.788	0.612	4.919	0.481
Time	1.100	1.100	0.932	0.000	1.571	0.471	1.494	0.394	0.300	0.300	0.936	0.000	0.917	0.617	0.789	0.489
Missingness model																
Intercept	-0.800	-0.867	0.954	0.067					-0.800	-0.905	0.956	0.105				
Time	1.250	1.224	0.960	0.026					1.250	1.205	0.942	0.045				
$y_i(\vartheta_c)$	-0.400	-0.410	0.920	0.010					-0.400	-0.407	0.940	0.007				
$r_{i-1}(\psi_c)$	2.500	2.600	0.950	0.100					2.500	2.612	0.948	0.112				
π	2.314	2.340	0.960	0.027					2.314	2.339	0.972	0.025				
Correlation structure																
σ_1	0.182	0.179	0.950	0.003					0.182	0.180	0.950	0.002				
σ_2	0.956	0.956	0.938	0.000					0.956	0.949	0.950	0.006				
σ_3	1.099	1.098	0.942	0.000					1.099	1.094	0.946	0.004				
$\rho_{1,2}$	0.847	0.854	0.942	0.007					0.847	0.846	0.954	0.001				
$\rho_{2,3}$	0.405	0.409	0.932	0.003					0.405	0.409	0.958	0.003				
								N=500								
Response Rate	Time 1	Time 2	Time 3						Time 1	Time 2	Time 3					
	0.911	0.805	0.587						0.910	0.666	0.427					
		Full Liklihood			GEE		WGEE			Full Liklihood			GEE		WGEE	
	TRUE	E.est	C.P	Bias	E.est	Bias	E.est	Bias	TRUE	E.est	C.P	Bias	E.est	Bias	E.est	Bias
Intercept	5.600	5.591	0.940	0.009	5.111	0.489	5.202	0.398	5.400	5.404	0.952	0.004	4.787	0.613	4.918	0.482
Time	1.300	1.309	0.934	0.009	1.778	0.478	1.698	0.398	0.300	0.295	0.954	0.005	0.917	0.617	0.788	0.488
Missingness model																
Intercept	-0.800	-0.907	0.932	0.107					-0.800	-0.872	0.920	0.072				
Time	1.250	1.164	0.938	0.086					1.250	1.226	0.940	0.024				
$y_i(\vartheta_c)$	-0.400	-0.403	0.946	0.003					-0.400	-0.410	0.962	0.010				
$r_{i-1}(\psi_c)$	2.500	2.606	0.952	0.106					2.500	2.600	0.964	0.100				
π	2.314	2.334	0.938	0.020					2.314	2.347	0.936	0.033				
Correlation structure																
σ_1	0.182	0.178	0.960	0.004					0.182	0.177	0.944	0.006				
σ_2	0.956	0.956	0.944	0.000					0.956	0.957	0.946	0.001				
σ_3	1.099	1.096	0.938	0.003					1.099	1.094	0.944	0.004				
$\rho_{1,2}$	0.847	0.850	0.958	0.002					0.847	0.845	0.948	0.003				
$\rho_{2,3}$	0.405	0.405	0.950	0.001					0.405	0.417	0.948	0.011				

[1] σ_i standard deviation of outcome at each time i. [2] $C.P$ coverage probability. [3] $E.est$ Mean estimation.

Table 2.7: Model comparison simulations, 1000 replicates

N=300										
Res.Rate time 1			0.9095				0.9102			
Res.Rate time 2			0.8395				0.8806			
Res.Rate time 3			0.6042				0.6368			
		Transition Model correctly specified		Troxel 1998 model			Transition Model		Troxel 1998 model correctly specified	
Parameter	TRUE	E.est	C.P	E.est	C.P	TRUE	E.est	C.P	E.est	C.P
Intercept	6.8	6.798	0.944	6.711	0.926	1.8	1.788	0.931	1.780	0.932
time	1.05	1.048	0.935	1.133	0.911	1.05	1.062	0.931	1.070	0.933
Missingness Model										
Intercept	-0.8	-0.948	0.954	0.721	0.575	-0.8	-1.115	0.941	-0.919	0.958
Time	1.25	1.145	0.943	2.397	0.852	1.25	1.022	0.944	1.190	0.949
$y_i(\vartheta_c)$	-0.4	-0.410	0.932	-0.289	0.785	-0.4	-0.407	0.950	-0.396	0.942
$r_{i-1}(\psi_c)$	2.5	2.674	0.963	-	-	0	0.229	0.959	-	-
π	0.91	0.910	1.000	0.910	1.000	0.91	0.914	1.000	0.916	1.000
Correlation										
σ_1	1.2	1.195	0.967	1.196	0.968	1.2	1.195	0.940	1.194	0.940
σ_2	2.6	2.593	0.948	2.549	0.919	2.6	2.594	0.936	2.586	0.936
σ_3	3	3.001	0.940	2.949	0.932	3	2.997	0.928	2.978	0.929
ρ_1	0.4	0.398	0.953	0.400	0.950	0.4	0.400	0.948	0.402	0.948
ρ_2	0.2	0.202	0.948	0.207	0.942	0.2	0.196	0.940	0.196	0.940

[1] Simulation sample size $n = 300$. replicates 1000. Res.Rate: Response Rate [2] σ_i standard deviation of outcome at each time i. [3] $C.P$ coverage probability. [4] $E.est$ Mean estimation.

Table 2.8: Simulation comparison study, 500 replicates

		AR(1) $\sigma^2\rho^{i-j}$		Exchangable $\sigma^2[\rho 1(i \neq j) + 1(i = j)]$		Toep(2) $\sigma^2_{\lvert i-j\rvert+1} 1(\lvert i-j\rvert < 2)$	
Response Rate time 1		0.910		0.910		0.910	
Response Rate time 2		0.844		0.844		0.844	
Response Rate time 3		0.614		0.616		0.610	
	TRUE	E.est	C.P	E.est	C.P	E.est	C.P
Intercept	6.8	6.809	0.954	6.799	0.954	6.809	0.932
Time	1.05	1.047	0.952	1.055	0.970	1.037	0.920
Missing Data Model							
Intercept	-0.8	-0.841	0.962	-0.934	0.960	-0.749	0.944
Time	1.25	1.238	0.970	1.123	0.954	1.331	0.966
$y_i(\vartheta_c)$	-0.4	-0.408	0.962	-0.397	0.952	-0.423	0.948
$r_{i-1}(\psi_c)$	2.5	2.581	0.966	2.593	0.968	2.622	0.948
π	0.904	0.911	1.000	0.911	1.000	0.910	1.000
correlation							
σ_{y1}	2.4	2.404	0.946	2.401	0.946	2.403	0.960
σ_{y2}	2.4	2.407	0.956	2.421	0.944	2.385	0.956
σ_{y3}	2.4	2.405	0.930	2.405	0.956	2.403	0.956
ρ_1	0.6	0.600	0.954	0.613	0.932	0.583	0.920
ρ_2	0.6	0.600	0.962	0.620	0.894	0.568	0.876

[1] Simulation sample size $n = 500$. replicates $R = 500$. [2] σ_i standard deviation of outcome at each time i.
[3] $C.P$ coverage probability. [4] $E.est$ Mean estimation.

Table 2.9: Non-normal data, 500 replicates

	Normal				Gamma			
Response Rate	0.910	0.855	0.613		0.910	0.855	0.609	
	TRUE	E.est	STD	C.P	TRUE	E.est	STD	C.P
Intercept	6.800	6.795	0.142	0.958	6.800	6.846	0.155	0.950
Time	1.050	1.054	0.065	0.958	1.050	1.000	0.080	0.912
Intercept	-0.800	-1.002	0.992	0.938	-0.800	-0.726	0.890	0.932
Time	1.250	1.055	1.063	0.946	1.250	1.417	0.988	0.916
$y_i(\vartheta_c)$	-0.400	-0.398	0.078	0.934	-0.400	-0.450	0.097	0.892
$r_{i-1}(\psi_c)$	2.500	2.662	0.805	0.972	2.500	2.760	0.640	0.972
π	2.314	2.337	0.205	0.930	2.314	2.336	0.205	0.956
Correlation structure								
σ_1	1.772	1.764	0.075	0.940	1.772	1.765	0.075	0.922
σ_2	1.887	1.898	0.083	0.962	1.887	1.957	0.088	0.846
σ_3	1.995	1.988	0.099	0.938	1.995	2.109	0.111	0.800
$\rho_{1,2}$	0.707	0.716	0.079	0.930	0.707	0.637	0.078	0.470
$\rho_{2,3}$	0.707	0.723	0.093	0.918	0.707	0.603	0.093	0.324
	Normal				Gamma			
Response Rate	0.909	0.857	0.609		0.910	0.856	0.607	
	TRUE	E.est	STD	C.P	TRUE	E.est	STD	C.P
Intercept	6.800	6.796	0.165	0.962	6.800	6.877	0.181	0.918
Time	1.050	1.056	0.086	0.970	1.050	0.967	0.106	0.866
Intercept	-0.800	-1.178	1.293	0.954	-0.800	-0.291	1.150	0.876
Time	1.250	0.876	1.395	0.942	1.250	1.964	1.332	0.876
$y_i(\vartheta_c)$	-0.400	-0.385	0.097	0.934	-0.400	-0.527	0.153	0.862
$r_{i-1}(\psi_c)$	2.500	2.724	1.010	0.988	2.500	2.868	0.696	0.982
π	2.314	2.323	0.204	0.944	2.314	2.338	0.205	0.932
Correlation structure								
σ_1	1.772	0.572	0.075	0.958	1.772	1.763	0.075	0.924
σ_2	1.887	0.635	0.086	0.958	1.887	1.966	0.094	0.840
σ_3	1.995	0.691	0.106	0.954	1.995	2.145	0.127	0.762
$\rho_{1,2}$	0.500	1.099	0.076	0.948	0.500	0.454	0.076	0.820
$\rho_{2,3}$	0.500	1.099	0.092	0.932	0.500	0.410	0.095	0.704
	Normal				Gamma			
Response Rate	0.910	0.857	0.600		0.910	0.857	0.600	
	TRUE	E.est	STD	C.P	TRUE	E.est	STD	C.P
Intercept	6.800	6.784	0.228	0.934	6.800	6.996	0.229	0.710
Time	1.050	1.064	0.161	0.892	1.050	0.858	0.159	0.498
Intercept	-0.800	-1.299	2.297	0.896	-0.800	1.267	2.702	0.538
Time	1.250	0.763	2.609	0.894	1.250	3.959	3.082	0.528
$y_i(\vartheta_c)$	-0.400	-0.392	0.234	0.876	-0.400	-0.820	0.294	0.490
$r_{i-1}(\psi_c)$	2.500	2.821	1.144	0.968	2.500	3.137	1.552	0.986
π	2.314	2.329	0.204	0.934	2.314	2.335	0.205	0.934
Correlation structure								
σ_1	1.772	1.767	0.076	0.958	1.772	1.761	0.075	0.934
σ_2	1.887	1.896	0.103	0.926	1.887	2.033	0.116	0.648
σ_3	1.995	2.000	0.144	0.940	1.995	2.294	0.182	0.562
$\rho_{1,2}$	0.000	-0.002	0.066	0.946	0.000	-0.004	0.064	0.942
$\rho_{2,3}$	0.000	0.004	0.081	0.932	0.000	-0.026	0.074	0.888

[1] Simulation sample size $n = 500$. replicates $R = 500$.
[2] σ_i standard deviation of outcome at each time i. [3] $C.P$ coverage probability.
[4] $E.est$ Mean estimation. STD Mean standard deviation

CHAPTER 3 : Pseudo-likelihood methods for transition models in longitudinal data with non-ignorable non-monotone missing data

3.1. Introduction

In a longitudinal study, subjects are observed as time progresses. A common problem is that repeated measurements are not fully observed due to missing responses or loss to follow up. Individuals can move in and out of the observed data set, giving rise to a large class of distinct "non-monotone" missingness patterns. The appropriate statistical methods differ according to the data structure and missing mechanism. When the missingness is MCAR (missing completely at random) or MAR (missing at random), data analysis is the most straightforward. Little and Rubin (1987) and Allison (2001) provide helpful terminology to describe missing data mechanisms and a comprehensive overview of potential methods. Most approaches can be categorized as selection models, pattern-mixture models or shared-parameter models depending on the factorization of the joint likelihood of the outcomes and missingness indicators. This chapter will focus on selection models.

Under the MCAR mechanism, the observed data can be viewed as a random subset of the complete data. With MAR data, the missingness mechanism depends only on observed quantities. Both mechanisms are termed "ignorable" if the parameters in the two parts of the model are distinct. Unbiased parameter estimates can be guaranteed using generalized estimating equations defined by Liang and Zeger (1986) when the missingness mechanism is MCAR, and using weighted estimating equations defined by Robins and Rotnitzky (1995) when the missingness mechanism is MAR. Neither method provides consistent unbiased estimators under informative dropout or non-ignorable (NI) missingess. The approaches to modeling longitudinal NI missing data depend on the data structure and type, variance/covariance structure, and proportion of missing data. Many proposed methods assume a multivariate Gaussian distribution for the outcomes, with different specifications of the covariance structure; these include (Verbyla and Cullis, 1990; Richard and Lynn,

1990; Munoz et al., 1992; Diggle and Kenward, 1994). However, all of these methods are full-likelihood methods, which require integration over the unobserved data. The parameter estimation has to be done numerically, and this can be computationally prohibitive, especially when the number of repeated assessments is large. Troxel et al. (1998b) proposed a pseudo-likelihood method (Parke, 1986; Gong and Samaniego, 1981) for analysis of continuous responses subject to non-ignorable non-monotone missing data. They treated the pairwise correlation coefficients ρ as nuisance parameters fixed at zero, which results in an independent likelihood over time. Specifically, their pseudo-likelihood assumed a simple Gaussian model for the outcome at each time, and also a marginal logistic regression model for the missingness probability at a given time that depends only on the possibly missing response at that time and the covariates, which are assumed to be fully observed. This pseudo-likelihood method significantly eases the computational complexities of the conventional likelihood-based method by reducing the high-dimensional integration to one-dimensional integration. This class of method could be viewed as an extension of composite marginal likelihood methods (Cox and Reid, 2004; Varin et al., 2011) which can be transferred into the non-ignorable non-monotone missing data framework.

Although this pseudo-likelihood method achieves asymptotically unbiased estimators of the regression parameters and missingness parameters if the model is correctly specified, these estimators can be highly inefficient due to the faulty assumption of independence of repeated measures across all measurement times. Parzen et al. (2007) proposed an alterative pseudo-likelihood method for binary data by using the joint distribution of all pairs of assessments to yield more efficient estimates. However, for $t = 1, 2, \ldots, T$ repeated observed times, there are a total of $\frac{T*(T-1)}{2}$ joint pairs to be calculated, which is still computationally impractical if T is large. Sinha et al. (2010) proposed a new bivariate pseudo-likelihood by counting all pairwise associations between the baseline response (first observation) and all subsequent responses. They assumed that baseline responses are always observed. However, the pairwise association becomes weak when the assessment is far from the baseline. In this article, we propose a new pseudo-likelihood that uses only adjacent pairs of observations.

The first-order Markov dependence structure has been shown to be a natural way to capture the correlation among repeated measurements in a longitudinal data framework. In practice, the AR(1) structure is often used to simplify the likelihood function. We show also that this method can be easily expanded to handle binary data.

The proposed methods were applied to data from the Penn Center of Excellence in Cancer Communication Research. Effective communication between patients and physicians is very important in cancer treatment and throughout the health care system. Effective exchange of information between patients, physicians, health care systems, and the surrounding environment determines how active participants are within the health care system. Many studies show a link between isolation and worse outcomes (Putt et al., 2009), including shorter survival time, worse quality of life, and lower rates of participation in recommended treatment programs. Patient adherence to treatment is normally higher in those who actively seek information about their treatment and quality of life from multiple channels (Tan et al., 2011). It is crucial to understand the relationship between patients, physicians, and the health care system, as well as the role of shared decision-making skills; how patients get, give, and discuss information and make health care decisions is important in cancer research, especially given the high demands that the health care system is facing.

The Effects of Public Information Study enrolled a total of 2010 patients diagnosed in 2005 with breast, colorectal, or prostate cancer selected from the Pennsylvania Cancer Registry. Subjects responded to at least one of three surveys, including 1520 patients who responded at wave 1, 1243 patients who responded at wave 2, and 1079 patients who responded at wave 3; these three survey occurred at yearly intervals beginning in fall 2006. The American Association for Public Opinion Research (AAPOR, 2006) response rates for the primary sample were 68%, 64%, and 61% for the respective cancer groups (Nagler et al., 2010); intermittent missingness patterns were common. Surveys were mailed to all participants using Dillman's design method (Dillman, 2010). All patients were first mailed an introductory letter explaining the purpose of the study and including instructions; the

surveys were mailed in a subsequent packet with a small monetary incentive ($3 or $5 for the short or long version of the survey). Reminder letters were sent after 2 weeks for subjects who did not return the survey. At follow-up assessments, contact was attempted with all patients, regardless of response to the prior year's survey. Patient consent was provided prior to participation, and the University of Pennsylvania Institutional Review Board reviewed and approved this study.

One of the study's research goals was to examine how the Patient-Clinician Information Engagement (PCIE) score affects cancer patients' attitudes and behaviors; in particular, researchers were interested in the amount of exercise the patients engaged in. For example, patients decide whether to increase exercise, to get radiation therapy, or to choose surgery after seeking and considering treatment information with their physicians; the decision making process may be influenced by both medical and non-medical information. The PCIE score was designed to measure these constructs using 8 items assessing whether, during the first few months of their cancer diagnosis, they had sought cancer, treatment, or quality-of-life information their own or other physicians or from other sources. Each of the 8 "Yes/No" questions was transformed to a Z-score, and the average of the 8 Z-scores formed the PCIE scale. We use the extent of exercise ("During an average week, how many days do you exercise?") as the primary outcome. The outcomes range from 0 to 7 by design; we treat these as continuous responses in this small interval. The Pearson correlation coefficients for the between-wave exerceise scores ranged from 0.58 to 0.64. We will assume that the correlation between each pair of adjacent assessments is the same.

The rest of this chapter is presented as follows. The proposed methods are described in Section 3.2, and illustrated with an analysis of the PCIE data in Section 3.3. A simulation study to address the performance of the methods is presented in Section 3.4. Section 3.5 provides a discussion and ideas for future work.

3.2. Model and Notation

Given a longitudinal data set, let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iT})'$ represent the vector of repeated measurements for subject i $(i = 1, \ldots, n)$ with T measurement times. Let $\mathbf{X}_i$ be a vector of p covariates observed on the ith subject. Because the repeated measurements are not fully observed at each time point $t = (1, \ldots, T)$, define a vector of missingness indicators $\mathbf{R}_i = (R_{i1}, R_{i2}, \ldots, R_{iT})$ to correspond with the outcome vector $\mathbf{Y}_i = (\mathbf{Y}_{i,obs}, \mathbf{Y}_{i,mis})$. Each element of R_i is defined as

$$R_{it} = \begin{cases} 0 & \text{if missing} \\ 1 & \text{if observed} \end{cases}$$

For each subject, the full data are given by the repeated measurements and missingness indicators with joint distribution $L(\theta, \beta | \mathbf{Y}_i, \mathbf{R}_i, \mathbf{X}_i) \propto P(\mathbf{Y}_i, \mathbf{R}_i | \mathbf{X}_i, \theta, \beta)$. By partitioning $\mathbf{Y}_i$ into $(\mathbf{Y}_{i,obs}, \mathbf{Y}_{i,mis})$, we can rewrite the joint likelihood in several ways. θ is parameter space associated with outcome process, and β is parameter space associated with missingness mechanism.

A selection model would specify the joint distribution using the marginal distribution of the repeated outcomes and the conditional distribution of missing indicators:

$$P(\mathbf{Y}_i, \mathbf{R}_i | \mathbf{X}_i, \theta, \beta) = P(\mathbf{Y}_{i,obs}, \mathbf{Y}_{i,mis} | \mathbf{X}_i, \beta) P(\mathbf{R}_i | \mathbf{Y}_{i,obs}, \mathbf{Y}_{i,mis}, \mathbf{X}_i, \theta).$$

We can assume that the complete data come from a multivariate normal distribution, $\mathbf{Y}_i \sim MVN(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where the mean structure $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \cdots \mu_{iT})$ depends on a p-dimensional covariate vector $\mathbf{X}_i$. We also assume a first-order Markov dependence structure for both the full outcome data and the missingness indicators, so that $f(Y_{it} | Y_{i1}, \ldots Y_{i,t-1}) = f(Y_{it} | Y_{i,t-1})$ and $f(R_{it} | R_{i1}, \ldots R_{i,t-1}) = f(R_{it} | R_{i,t-1})$. Since our proposed pseudo-likelihood uses only adjacent pairs of observations, we will let $\sigma^2 = var(Y_{it})$ and $\rho = corr(Y_{it}, Y_{i,t+1})$. Then we

can denote the conditional likelihood as

$$Y_{it}|Y_{i,t-1} \sim N\Big\{\mu_{it} + \rho(Y_{i,t-1} - \mu_{i,t-1}), \sigma^2(1-\rho^2)\Big\}.$$

3.2.1. Missingness mechanism model

Unlike other approaches to modeling the missingness mechanism, we are interested in the transition probability of the missingness indicators R_{it} given R_{it-1}. Conditional on each time t, the missingness mechanism becomes a two-state Markov chain. We model the transition probabilities $\pi_{jk} = Pr(R_{it} = j|R_{i,t-1} = k, Y_{it}, X_{it})$, $j = 0, 1; k = 0, 1$ as

$$\begin{pmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{pmatrix}$$

which satisfy the equation $\pi_{00} + \pi_{01} = \pi_{10} + \pi_{11} = 1$. We assume that the initial state is independent, and define n_{ijk} as the number of times in the whole sequence that k is followed by j:

$$\begin{aligned} n_{ijk} &= \sum_{t=1}^{T} I(R_{it} = j|R_{i,t-1} = k) \\ n_{ij.} &= \sum_{k} n_{j,k}, \quad n_{i.k} = \sum_{j} n_{j,k}. \end{aligned}$$

Then the missingness mechanism can be written as

$$\begin{aligned} \mathbb{L}_i &= \pi_{00}^{n_{i00}} \pi_{01}^{n_{i01}} \pi_{10}^{n_{i10}} \pi_{11}^{n_{i11}} \\ &= \prod_{t=2}^{T} \prod_{j=0}^{1} \prod_{k=0}^{1} \pi_{jk}(t)^{I(R_{it}=j|R_{i,t-1}=k)}. \end{aligned}$$

This becomes a product of binomial distributions. To avoid the unstable estimation problems for binary outcomes near the boundary of the parameter space, we estimate the probability of missingness at each time t using a joint beta-binomial distribution instead of

traditional estimation using logistic regression. From a Bayesian perspective, the beta distribution is the conjugate prior distribution for the parameters of the binomial distribution. The parameters of the beta distribution can be viewed as pseudo-counts of "response" and "non-response" to be added to the observed counts of responses and non-responses.

Given time $t-1$, the missingness mechanism follows $(R_{it}|R_{i,t-1}=k) \sim Bernoulli(\pi_{ikt})$; we impose a beta distribution on the missingness probability, $\pi_{ikt} \sim Beta(a_{ikt}, b_{ikt})$ Then we have

$$\begin{aligned} f(R_{it}|R_{i,t-1}, y_{it}, \pi) &= \prod_{k=0}^{1} \pi_{k1}^{I(R_{it}=1)I(R_{i,t-1}=k)}(1-\pi_{k1})^{[1-I(R_{it}=1)]I(R_{i,t-1}=k)} \\ f(\pi_{k1}|a_{k1}, b_{k1}) &= \frac{\Gamma(a_{k1}+b_{k1})}{\Gamma(a_{k1})\Gamma(b_{k1})} \times \pi_{k1}^{a_{k1}-1}(1-\pi_{k1})^{b_{k1}-1}. \end{aligned}$$

Integrating the π out, the mixture function can be expressed as

$$\begin{aligned} & f(R_{it}|R_{i,t-1}, a_{ik1}, b_{ik1}, y_{it}) = \int_0^1 f(R_{it}|R_{i,t-1}=k, \pi_{ikt}) f(\pi_{ikt}|a_{ikt}, b_{ikt}, y_{it}) d\pi_{ikt} \\ = & \prod_{k=0}^{1} \frac{\Gamma(a_{ik1}+b_{ik1})}{\Gamma(a_{ik1})\Gamma(b_{ik1})} \\ \times & \frac{\Gamma(a_{ik1}+I(R_{it}=1)I(R_{i,t-1}=k))\Gamma(b_{ik1}+[1-I(R_{it}=1)]I(R_{i,t-1}=k))}{\Gamma(a_{ik1}+b_{ik1}+I(R_{i,t-1}=k))} \end{aligned}$$

with $a_{ik1} = \exp(\zeta_1 \mathbf{X_{it}} + \vartheta_\mathbf{1} \mathbf{Y_{it}} + \psi_\mathbf{1} \mathbf{R_{i,t-1}})$ and $b_{ik1} = \exp(\zeta_2 \mathbf{X_{it}} + \vartheta_\mathbf{2} \mathbf{Y_{it}} + \psi_\mathbf{2} \mathbf{R_{i,t-1}})$. However, the link function chosen could be different, resulting in a different missingness mechanism model. For given $R_{i,t-1}=0$, the transition probability can be denoted as

$$P(R_{it}=l|R_{i,t-1}=0, Y_{it}, X_{it}) = \begin{cases} \frac{1}{1+\exp((\zeta_1-\zeta_2)\mathbf{X_{it}}+(\vartheta_\mathbf{1}-\vartheta_\mathbf{2})\mathbf{Y_{it}})} & \text{if} \quad l=1 \\ \frac{1}{1+\exp(-(\zeta_1-\zeta_2)\mathbf{X_{it}}-(\vartheta_\mathbf{1}-\vartheta_\mathbf{2})\mathbf{Y_{it}})} & \text{if} \quad l=0 \end{cases}$$

For given $R_{i,t-1}=1$,

$$P(R_{it}=l|R_{i,t-1}=1, Y_{it}, X_{it}) = \begin{cases} \frac{1}{1+\exp((\zeta_1-\zeta_2)\mathbf{X_{it}}+(\vartheta_\mathbf{1}-\vartheta_\mathbf{2})\mathbf{Y_{it}}+(\psi_\mathbf{1}-\psi_\mathbf{2}))} & \text{if} \quad l=1 \\ \frac{1}{1+\exp(-(\zeta_1-\zeta_2)\mathbf{X_{it}}-(\vartheta_\mathbf{1}-\vartheta_\mathbf{2})\mathbf{Y_{it}}-(\psi_\mathbf{1}-\psi_\mathbf{2}))} & \text{if} \quad l=0 \end{cases}$$

Notice that if $\vartheta_1 - \vartheta_2 \neq 0$ then the missingness mechanism is indeed non-ignorable since the probability of missingness depends on the unobserved outcome Y_{it}. In practice, only the difference of the parameters is identifiable, not the individual parameters. We let $\zeta_c = \zeta_1 - \zeta_2$, $\vartheta_c = \vartheta_1 - \vartheta_2$ and $\psi_c = \psi_1 - \psi_2$ be the final parameters in the missingness mechanism model, where ζ_c is the coefficient of the covariates, ϑ_c is the coefficient of the current observed outcome y_{it}, and ψ_c is the coefficient of the previous missingness indicator $r_{i,t-1}$. Note that the covariates $\mathbf{X_{it}}$ can include effects for time and/or interaction terms between time and other variables, making it highly flexible and able to accommodate a wide range of effects on the missing data probabilities.

The link function for parameters a_{ikt} and b_{ikt} of the beta distribution could be chosen as something other than a simple exponential function, and this will result in a different missingness mechanism model. The missingness mechanism model could be expanded similarly to a Dirichlet-multinomial distribution from the current beta-binomial distribution when modeling more than two missingness states, such as "observed," "intermediate missing," and "drop out."

3.2.2. Independent Pseudo-likelihood (IPL)

Troxel et al. (1998b) proposed a pseudo-likelihood approach for the analysis of continuous longitudinal responses subject to non-ignorable non-monotone missing data. They treated the pairwise correlation coefficients ρ as nuisance parameters fixed at zero. Then they modeled the repeated measurements independently over time after applying this working independence assumption. To describe this pseudo-likelihood method, let $f(y_{it}, r_{it}|x_i, \theta, \beta)$ be the marginal distribution of (Y_{it}, R_{it}) at each time t; then the observed pseudo-likelihood can be denoted:

$$\mathcal{L}(\theta, \beta)^{obs} = \prod_{i=1}^{N} \prod_{t=1}^{T} f(y_{it}, r_{it}|x_i, \theta, \beta)^{r_{it}} \left[\int_{y_{it}} f(y_{it}, r_{it}|x_i, \theta, \beta) dy_{it} \right]^{1-r_{it}}.$$

Further, let $f(y_{it}|x_i, \theta)$ be a normal distribution with $N(\mu_{it}, \sigma^2)$, and $f(r_{it}|y_{it}, x_i, \beta)$ is

Bernoulli distribution with probability of being observed. The above formula can be expanded as:

$$\begin{aligned}
\mathcal{L}(\theta,\beta)^{obs} &= \prod_{i=1}^{N}\prod_{t=1}^{T} f(y_{it},r_{it}|x_i,\theta,\beta)^{r_{it}}\left[\int_{y_{it}} f(y_{it},r_{it}|x_i,\theta,\beta)dy_{it}\right]^{1-r_{it}} \\
&= \prod_{i=1}^{N}\prod_{t=1}^{T}\{f(y_{it}|x_i,\theta)f(r_{it}|y_{it},x_i,\beta)\}^{r_{it}} \\
&\times \left[\int_{y_{it}} f(y_{it}|x_i,\theta)f(r_{it}|y_{it},x_i,\beta)dy_{it}\right]^{1-r_{it}} \\
&= \prod_{i=1}^{N}\prod_{t=1}^{T}\{f(y_{it}|x_i,\theta)\pi_{it}\}^{r_{it}} \\
&\times \left[\int_{y_{it}} f(y_{it}|x_i,\theta)(1-\pi_{it})dy_{it}\right]^{1-r_{it}}
\end{aligned}$$

The estimators $(\hat{\theta},\hat{\beta})$ can be obtained by setting the log pseudo-score vector equal to zero and solving:

$$\mathcal{S}(\theta,\beta) = \frac{\partial}{\partial(\theta,\beta)}\log\mathcal{L}(\theta,\beta)^{obs}.$$

Using method of moments ideas, this estimator $(\hat{\theta},\hat{\beta})$ is consistent since it can be shown that $E[\mathcal{S}(\theta,\beta)] = 0$, and the estimator $(\hat{\theta},\hat{\beta})$ is consistent for the true parameters θ,β (Troxel et al., 1998b).

Although this pseudo-likelihood method achieves asymptotically unbiased estimators of the regression parameters and missingness parameters if the model is correctly specified, these estimators can be highly inefficient due to the faulty assumption of independence of repeated measures across all measurement times, and this pseudo-likelihood method ignores the covariance structure entirely.

3.2.3. Proposed Transition Pseudo-likelihood (TPL)

Instead of focusing on the marginal likelihood, we consider a pseudo likelihood based on the conditional density of all adjacent pairs. Let J be the index set, $J = \big(\{1,2\},\{2,3\},\cdots,\{T-$

$2, T-1\}, \{T-1, T\}\big)$. To avoid the computational burden of the full likelihood approach and take advantage of the pseudo-likelihood framework, we propose to use only $T-1$ adjacent pairs. A first-order Markov dependence structure is assumed for longitudinal data. We re-write the joint density function as:

$$\begin{aligned}
\mathcal{PL}_i &= \prod_{i=1}^{n} p(y_{i1}, r_{i1}|x_i, \theta, \beta) \prod_{j=2}^{t} p(y_{ij}, r_{ij}|y_{i,j-1}, r_{i,j-1}, x_i, \theta, \beta) \\
&= \prod_{i=1}^{n} p(y_{i1}|x_i, \theta) p(r_{i1}|y_{i1}, x_i, \beta) \prod_{j=2}^{t} p(y_{ij}|y_{i,j-1}, x_i, \theta) p(r_{ij}|r_{i,j-1}, y_{ij}, y_{i,j-1}, x_i, \beta) \\
&= \prod_{i=1}^{n} p(y_{i1}|x_i, \theta) p(r_{i1}|y_{i1}, x_i, \beta) \prod_{j=2}^{t} p(y_{ij}|y_{i,j-1}, x_i, \theta) p(r_{ij}|r_{i,j-1}, y_{ij}, x_i, \beta)
\end{aligned}$$

The proposed transitional pseudo-likelihood (TPL) method is to assume independence of each time within subject and $f(y_{it}, r_{it}|y_{i,t-1}, r_{i,t-1}) \perp f(y_{i,t-1}, r_{i,t-1}|y_{i,t-2}, r_{i,t-2})$. The observed pseudo-likelihood function is denoted as

$$\begin{aligned}
\mathcal{PL}_i^{obs} &= L_0 * L_1 * L_2 * L_3 * L_4 \\
L_0 &= \prod_{i=1}^{n} \Big[p(y_{i1}) p(r_{i1}|y_{i1}) \Big]^{r_{i1}} * \Big[\int_{y_{i1}} p(y_{i1}) p(r_{i1}|y_{i1}) dy_{i1} \Big]^{1-r_{i1}} \\
L_1 &= \prod_{j=2}^{t} \Big[p(y_{ij}|y_{i,j-1}) p(r_{ij}|r_{i,j-1}, y_{i,j}) \Big]^{r_{ij} r_{i,j-1}} \\
L_2 &= \Big[\int_{y_{i,j-1}} p(y_{ij}|y_{i,j-1}) p(r_{ij}|r_{i,j-1}, y_{ij}) p(y_{ij-1}) dy_{ij-1} \Big]^{(1-r_{i,j-1}) r_{ij}} \\
L_3 &= \Big[\int_{y_{ij}} p(y_{ij}|y_{i,j-1}) p(r_{ij}|r_{i,j-1}, y_{i,j}) dy_{ij} \Big]^{r_{i,j-1}(1-r_{ij})} \\
L_4 &= \Big[\iint_{y_{i,j-1}, y_{ij}} p(y_{ij}|y_{i,j-1}) p(r_{ij}|r_{i,j-1}, y_{i,j}) p(y_{ij-1}) dy_{ij-1} dy_{i,j} \Big]^{(1-r_{i,j-1})(1-r_{i,j})}.
\end{aligned}$$

For the conditional distribution of $f(y_{it}|y_{i,t-1})$, we impose the prior density of $y_{i,t-1}$ to integrate out time point $t-1$ if the data in $t-1$ is unobserved. This situation occurs in the setting denoted as L_2 and L_4 above. Therefor we consider the $y_{i,t-1}$ as a random variable if the assessment is unobserved.

The proposed TPL is for continuous outcomes; however, this can be easily extended for binary outcomes as:

$$
\begin{aligned}
\mathcal{PL}_i^{obs} &= L_0 * L_1 * L_2 * L_3 * L_4 \\
L_0 &= \prod_{i=1}^{n} \Big[p(y_{i1}) p(r_{i1}|y_{i1}) \Big]^{r_{i1}} * \Big[\sum_{y_{i1}} p(y_{i1}) p(r_{i1}|y_{i1}) \Big]^{1-r_{i1}} \\
L_1 &= \prod_{j=2}^{t} \Big[p(y_{ij}|y_{i,j-1}) p(r_{ij}|r_{i,j-1}, y_{i,j}) \Big]^{r_{i,j-1} r_{ij}} \\
L_2 &= \Big[\sum_{y_{i,j-1}} p(y_{ij}|y_{i,j-1}) p(r_{ij}|r_{i,j-1}, y_{i,j}) p(y_{ij-1}) \Big]^{(1-r_{ij-1}) r_{ij}} \\
L_3 &= \Big[\sum_{y_{ij}} p(y_{ij}|y_{i,j-1}) p(r_{ij}|r_{i,j-1}, y_{i,j}) \Big]^{r_{i,j-1}(1-r_{ij})} \\
L_4 &= \Big[\sum \sum_{y_{i,j-1}, y_{ij}} p(y_{ij}|y_{i,j-1}) p(r_{ij}|r_{i,j-1}, y_{i,j}) p(y_{i,j-1}) \Big]^{(1-r_{ij-1})(1-r_{i,j})}.
\end{aligned}
$$

The pseudo-score function is defined as:

$$
\mathcal{S}_T(\theta, \beta) = \sum_{i=1}^{n} \mathcal{S}_{Ti}(\theta, \beta) = \frac{\partial}{\partial(\theta, \beta)} \log \mathcal{PL}_i^{obs},
$$

and the maximum pseudo-likelihood estimate is the solution to $\mathcal{S}_T(\hat{\theta}, \hat{\beta}) = \mathbf{0}$. Heuristically, using method of moments ideas, the transition pseudo-score estimator is consistent if the distributions $f(y_{it}, y_{i,t-1}, r_{it}, r_{i,t-1}|\mathbf{X_i}, \theta, \beta)$ are correctly specified. Troxel et al. (1998b) gave proof of the consistency of the pseudo-likelihood estimator. It can be shown that $\mathbf{E}[\mathcal{S}_T(\theta, \beta)] = 0$ at the true (θ, β). In practice, we obtain $(\hat{\theta}, \hat{\beta})$ by maximizing the log-pseudolikelihood directly, but the solution satisfies $\mathcal{S}_T(\hat{\theta}, \hat{\beta}) = \mathbf{0}$. The variance must be adjusted to obtain correct inference because of the faulty independence assumption. We accomplish this with the commonly-used sandwich estimator as in Liang and Zeger (1986):

$$
\Sigma = \Big[\frac{1}{n} E\Big\{ \frac{\partial \mathcal{S}_T(\theta, \beta)}{\partial(\theta, \beta)} \Big\} \Big]^{-1} \frac{1}{n} \sum_{i=1}^{n} E\{ \mathcal{S}_{Ti}(\theta, \beta) \mathcal{S}'_{Ti}(\theta, \beta) \} \Big[\frac{1}{n} E\Big\{ \frac{\partial \mathcal{S}_T(\theta, \beta)}{\partial(\theta, \beta)} \Big\} \Big]^{-1}.
$$

Furthermore, the variance estimate $\hat{\Sigma}$, is obtained by simply replacing (θ, β) by $(\hat{\theta}, \hat{\beta})$ in

the above expression.

The Gauss-Hermite quadrature rule will be used to approximate the integration since there is no closed form for the observed pseudo-likelihood function. For any given function $f(t)$ we can approximate an integral as a summation following the transformation of Liu and Pierce (1994):

$$\int_{\infty}^{\infty} f(t)\phi(t;\mu,\sigma)dt \simeq \sum_{i=1}^{m} \frac{w_i}{\sqrt{\pi}} f(\mu + \sqrt{2}\sigma\tau_k).$$

The above expression will be exact if the given function $f(t)$ is polynomial with degree less than $2m-1$. $\phi(t;\mu,\sigma)$ is a normal density with mean μ and standard deviation σ. (w_k, τ_k) is the kth Gauss-Hermite weight and nodes (zero root of the m^{th} order Hermite polynomials).

3.3. Example: Analysis of PCIE Data

3.3.1. Application to PCIE Data sets

Survey studies increasingly focus on questionnaires from patients with different health issues, stages of disease, types of cancer; both medical and more general non-medical information is needed for health providers and decision makers to better understand the behavior changes of subjects. Intuitively, patient behaviors involving attitude change, information seeking, and willingness to respond to questionnaires are related to health status. It is reasonable to expect that patients may be less likely to respond due to worsened health status; this may be a function of different kinds of health information including disease type, patient self-efficacy, and their surrounding environment, and may contribute to informative missingness.

Table 3.1 lists the missingness patterns in PCIE data; all eight possible patterns are represented in the data, including non-monotone patterns. In practice, pattern 1, which has missing data at all three waves, carries no information and will be excluded from the study. We use the extent of exercise ("During an average week, how many days do you exercise?") as the primary outcome. The outcome ranges from 0 to 7; we treat these as continuous re-

sponses in this small interval. There were 85.66% of patients who responded to the baseline survey, 61.75% who returned the survey in wave 2, and 56.03% who answered the questions in wave 3; response rates for specific variables are given in Table 3.2. We calculated the Pearson correlation coefficients for the exercise outcomes, which shows that the correlation is 0.644 between waves 1 and 2, and 0.618 between waves 2 and 3; the correlation between waves 1 and 3 is somewhat smaller at 0.585. We use an AR(1) correlation in the data analysis and simulation sections since our proposed model only incorporates the correlation of adjacent pairs of assessments.

Table 3.3 lists all patient characteristics of interest for both the marginal model and the missingness model. There are a total of 2010 cancer patients who responded to at least one of the three surveys, including 1520 patients who responded at wave 1, 1243 patients who responded at wave 2, and 1079 patients who responded at wave 3; these three surveys occurred at yearly intervals. The cohort includes both male and female patients with cancer stage from mild (stage 0) to severe (stage 4). Age at cancer diagnosis ranged from 23 to 103. The PCIE score was measured using 8 items as described earlier, indicating whether they had 1) sought information about treatment from their treating physician; 2) sought treatment information from other physicians or health professionals; 3) actively looked for information about their cancer from their treating physician; 4) looked for information about their cancer from other physicians or health professionals; 5) discussed information from other sources with their treating physician; 6) received suggestions from their treating physician to get information from other sources; 7) actively looked for information about quality of life issues from their treating physician; and 8) looked for quality of life information from other physicians or health professionals. Each of the 8 "Yes/No" questions was answered and was transformed to a Z-score, and the average of the 8 Z-scores formed the PCIE scale. The summary in Table 3 provides the variation in the PCIE score at each assessment time. Clearly patients with colorectal cancer were more likely to respond at wave 1 and less likely at wave 3. Breast cancer and prostate cancer patients showed the opposite pattern. Patients with severe cancer stage were less likely to respond to the survey

at wave 3 compared with wave 1.

The parameters are estimated using the proposed method and compared to full-likelihood method, independent pseudo-likelihood, a GEE model which assumes MCAR missingness and a weighted GEE (WGEE) which assumes MAR. Both GEE and WGEE can be treated as "ignorable" mechanisms. The original PCIE study analysis considered the missingness mechanism as either MAR or MCAR which may have resulted in biased estimates. Our WGEE approach is a modification of the standard GEE to address missingness in the data. The response probabilities are first calculated using a logistic model for the missingness indicators; the inverse of these estimated probabilities form the corresponding weights. The missingness mechanism model used "cancer type," "gender," "age at diagnosis," "cancer severity," "PCIE score," and the previous missingness indicator to predict current missingness indicators. For "ignorable" data, WGEE will produce unbiased estimators if the underlying data is MAR or MCAR. GEE may have biased estimates if the underlying data are MAR.

Because missing covariate data was not of primary interest, a multiple imputation method was used to complete the missing covariates. Rubin (1987) proposed a multiple imputation method using a Monte Carlo approach in which the missing values are replaced by $m > 1$ simulated versions. We generated $m = 20$ replicates in our study. Each of the imputed datasets is analyzed using the proposed method, the full-likelihood method, the independent pseudo-likelihood, the GEE model, and the weighted GEE model. The combined parameter estimates and confidence intervals from the $m = 20$ data sets follow Rubin (1987)'s multiple imputation rule.

Table 3.4, we list the parameter estimates after combined 20-fold imputation. The coefficient for Y_{it} indicates whether the missingness mechanism depends on the potentially unobserved outcome; a test of this parameter represents a test for non-ignorability. The coefficient for $R_{i,t-1}$ indicates whether the previous response has an effect on the likelihood of response at the current assessment; $R_{i,t-1} = 1$ indicates that the previous response was observed.

Clearly, there is a statistically significant effect in the missingness model for the coefficients of both Y_{it} $[-0.113\ (-0.188, -0.037)]$ and $R_{i,t-1}$ $[-0.802\ (-0.938, -0.666)]$ in the TPL model, which indicates that the MCAR assumption is invalid. The coefficients of Y_{it} and $R_{i,t-1}$ are both negative, indicating a negative relationship with the missingness indicator. That is, patients who exercise more days per week are less likely to have missing survey responses. They also tend to return the questionnaire if they have responded to previous one. The full-likelihood model shows the same trend as well, and the IPL method is less efficient for testing the coefficient of Y_{it}, since the correlation is large ($\rho = 0.629$). The coefficients for the other covariates indicate that patients who have prostate cancer $[-0.304\ (-0.623, 0.015)]$ are not statistically significantly different compared with the significant result from the full-likelihood method $[-0.301\ (-0.595, -0.007)]$; this may suggest loss of efficiency with the TPL model. Unsurprisingly, patients with advanced cancer (stage 4) $[0.726\ (0.191, 1.261)]$ are more likely to have a missing response. "Wave" has coefficient $[0.283\ (0.161, 0.405)]$ which suggests that patients tend to be less responsive to the survey as time passes; this is typical in repeated measures studies.

The marginal estimates from our proposed outcome model for exercise are somewhat larger than the ones from either the GEE or WGEE approach. However, the significance levels are consistent across the models. Only "age at diagnosis" and "cancer stage" are statistically significant. "Age at diagnosis" has coefficient $0.010\ (0.003, 0.018)$, indicating that older patients engage in more exercise then younger patients. The coefficient of "cancer stage" $[-0.549\ (-1.059, -0.040)]$ indicates a negative correlation with outcome. Patients tend to reduce the amount of exercise when their cancer becomes more severe. PCIE did not show a statistically significant effect in four of the models, with the exception of the IPL model, suggesting that in this sample, patient health behaviors are not significantly affected by patient engagement with the health care system as measured by the PCIE score. We assessed interaction terms in this model as well, to check whether the relationship between PCIE score and exercise might be changing over time, but found no evidence for this.

Although the MCAR and MAR assumption is apparently invalid, both GEE and WGEE models show similar trends to the proposed model; while most of the parameters estimates are attenuated, the inferential conclusions are unchanged in this example. The weighted GEE approach provides similar results to the GEE in this example, which may be due in part to the large sample size. The coefficients from the independent pseudo-likelihood are mostly consistent with results from the full-likelihood method and the transition pseudo-likelihood method; however there are some small departures due to the high correlation discussed in Troxel et al. (1998b), as expected.

3.4. Simulation Study

3.4.1. Simulation results

In this section we use a simulation study to assess model performance in small samples, addressing the basic issues of bias in the parameter estimates and computing coverage probabilities. We simulated N subjects with three potential measurement times, for $N = 300$, $N = 500$ and $N = 1000$. We compared the proposed model (TPL) and the independent pseudo-likelihood model (IPL) using the correct non-ignorable missingness mechanism in Table 3.5. There is about 10% missing at time 1, 25% at time 2, and around 35% missing at time 3. The true parameters were selected to generate the same proportion of missing data across multiple scenarios. The correctly specified original model for the IPL will become a misspecified model since the coefficient of the previous missing indicator is not zero (e.g., $\psi_c = 1.2$). We restricted the number of occasions to $T = 3$ and consider a simple two-group study design configuration with time interaction. However, the proposed TPL method has the same computational requirements as IPL, which can be used in studies with many repeated assessments.

Let $x_j = 0, 1$ indicate treatment group. The continuous outcomes, denoted by (Y_{i1}, Y_{i2}, Y_{i3}),

are assumed to follow a multivariate normal distribution, with joint probabilities

$$f(\mathbf{Y}) = (2\pi)^{-N/2} \det(\Sigma)^{-1/2} \exp\big(-\frac{1}{2}(\mathbf{Y}-\mu)^T \Sigma^{-1} (\mathbf{Y}-\mu)\big)$$

where $\mu_{it} = \alpha_0 + \alpha_1 t + \alpha_2 x_j + \alpha_3 t * x_j$, $t = 1, 2, 3$. For the simulation study, we choose $\alpha_0 = 11.5$, $\alpha_1 = 1.05$, $\alpha_2 = 0.25$, $\alpha_3 = 0.21$. The correlation structure is TOEP(2) in order to make the simulated covariance structure as close as possible to model assumptions. Standard deviation σ and pairwise correlation ρ was set at $\rho_{t,t-1} = 0.4$ and $\sigma = 1.2$. A variety of different correlation structures were examined and the same overall pattern of results was obtained in Table 3.6. The following true non-ignorable missingness mechanism was applied:

$$logit[pr(R_{it} = 1 | r_{i,t-1}, y_{it}, x_i, z_i)] = \beta_0 + \beta_1 z_i + \beta_2 x_i + \beta_3 t + \beta_4 t * x_i + \beta_5 y_{it} + \beta_6 r_{i,t-1}$$

where z_i is a covariate only for the missingness model.

In the simulations reported in Table 3.5, both models show approximately unbiased estimation of marginal parameters. TPL has consistently higher efficiency than IPL model. The coverage probabilities are close to each other and the bias is small. The correctly specified model for TPL becomes a miss-specified model for IPL method which is reflected in the estimation of the missingness model. The bias becomes large for the parameters in the missingness model in IPL model, and the coverage probability drops. Both methods can have approximately unbiased estimation in variance, but only TPL can provide an estimate of the correlation.

The correctly specified original model from IPL will become a mis-specified model if the coefficient of the previous missing indicator is not zero. Our proposed model will be over-specified if the parameter of the previous missing indicator is zero. Table 3.6 shows these comparison results. When the parameter (ψ_c) of the previous missing indicator is not zero, the estimates from our TPL model are unbiased and have high coverage probabilities. The

IPL model has good estimation in the marginal model and variance-covariance structure, but poor estimation in the missingness model. This is not surprising, since the missingness model is miss-specified. When the parameter (ψ_c) of the previous missing indicator is zero, both models have very good estimation. The proposed model uses a small value to estimate the ψ_c with 95% coverage rate including zero.

Next, we fit the proposed model with three different covariance structures to see how our model handles a miss-specified correlation matrix in Table 3.7. Our transition model uses only adjacent pairs of observations. We assume all correlation between each adjacent pairs of assessment are same, and assume same the variance over time to simplify the simulation. In practice the AR(1) (first-order autoregressive) structure is widely used, with covariance $\sigma^2\rho^{i-j}$ for the $(i,j)th$ element. There are only two parameters needed. Another two correlation structures used for comparison are the exchangeable ($\sigma^2[\rho 1(i \neq j) + 1(i = j)]$) structure and the TOEP(2) (Banded Toeplitz $\sigma^2_{|i-j|+1}1(|i-j| < 2)$) structure, which is as close as possible to our model assumptions.

The AR(1), exchangeable, and TOEP(2) structure for $T = 3$ are written respectively as:

$$\Sigma = \left\{ \begin{array}{ccc} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 \end{array} \right\}_{AR(1)} ; \Sigma = \left\{ \begin{array}{ccc} \sigma^2 & \sigma^2\rho & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2\rho & \sigma^2 \end{array} \right\}_{Exch} ; \Sigma = \left\{ \begin{array}{ccc} \sigma^2 & \sigma_1 & 0 \\ \sigma_1 & \sigma^2 & \sigma_1 \\ 0 & \sigma_1 & \sigma^2 \end{array} \right\}_{TOEP(2)}$$

The comparison table is listed in Table 3.7. The proposed model can handle the TOEP(2) structure, as well as exchangeable and AR(1). Our model performs quite well in estimating the marginal effects and missingness coefficients for a mis-specified correlation matrix. The variances are estimated with high coverage probabilities. Simulations show that our proposed method is robust to specification of the correlation matrix. However, the TPL model can provide a estimation of correlation structure while IPL cannot.

3.4.2. Non-normal data

In this section we compare the proposed models in cases with different underlying assumptions about the true data distribution. We simulated two data sets with the same expected values but with different distributions. One scenario was simulated from a trivariate normal distribution. A second scenario was simulated using a trivariate gamma distribution. A Clayton copula, which is an asymmetric Archimedean copula, was used to generate the trivariate gamma data. The dependence structure of the trivariate gamma followed an exchangeable correlation structure. We used Kendall (1976)'s formula to assure the same covariance structure between the trivariate normal and trivariate gamma data. We generated three correlation structures, with high ($\rho = 0.707$), low ($\rho = 0.5$) and zero ($\rho = 0$) pairwise correlation, with sample size $n = 300$ replications to examine the model's performance. We let the mean $\mu_{it} = \alpha_0 + \alpha_1 t$, $t = 1, 2, 3$ for both normal and gamma data. The variance was calculated through the Clayton copula to match the normal distribution data ($\sigma_1 = 2.145, \sigma_2 = 2.241, \sigma_3 = 2.332$). Both our proposed model and the IPL model assume the same variances over time. The comparison table is listed in Table 3.8. Our proposed model performed quite well even with the mis-specified distribution compared as the IPL model. The estimator becomes less efficient when the assesments are highly correlated ($\rho = 0.707$); this is not surprising since in this scenario the variance-covariance structure departs more drastically from the assumed structure. The marginal effects and missingness model are still estimated well when the underlying data distribution is not normal. The correlation coefficients are estimated well; however, the variance is estimated poorly.

3.5. Discussion

We have presented an extension of the pseudo-likelihood method to handle non-monotone and non-ignorable missing data. We assume a first-order Markov structure in both the complete data and missingness mechanism, which is a natural way to capture the correlation among repeated measurements in a longitudinal data framework. The estimation of marginal effects is generally robust to correct specification of the covariance matrix. Be-

cause of the assumptions inherent in the models, the broad range of possible missing data configurations and underlying probability distributions generating the data, it is difficult to draw general conclusions from the limited simulation study. However, based on our simulation study, we have shown that our proposed TPL model can handle longitudinal data with various covariance structures. Our proposed TPL model is no more computationally intensive than the IPL model, which makes this model easily used in situations with a large number of assessments.

Our transition model can be easily extended to model more than two states such as dropout or intermittent missingness. The numerical integration provides an accurate approximation but at the cost of increased computational complexity. We occasionally encountered a multimodal likelihood surface in our study. A method to handle such as surface is to choose a vector of starting values by using GEE estimates to get the starting point as close as the true values as possible. There are too many classes of correlation structure to explore them all; however, the proposed model can handle a mis-specified correlation to some extent. In simulation studies with a variety of miss-specified correlation structures, the marginal effects and missingness effects consistently have high coverage probabilities as long as the correlation among pairs is nonzero.

Given the increasing interest in health care reform and structural changes in health care systems, more and more survey studies are being designed to better understand the relationships among patients, physicians, and the broader health care system. In many such studies, however, sample sizes are limited based on the disease under study, the geographic area, and the availability of medical information. Small sample sizes with a large proportion of missing information become a vexing problem for researchers trying to evaluate the associations of interest. The missingness probability is often related to the very outcomes under study, e.g., when patients fail to respond because of worse health outcomes. In the example studied here, level of exercise may well serve as a proxy for general health status; ignoring this information in the analysis can lead to seriously biased results.

Table 3.1: Missingness patterns in PCIE study

Pattern			Number of cases	Pattern			Number of cases
0	0	0	166	1	0	0	457
0	0	1	26	1	0	1	118
0	1	0	77	1	1	0	231
0	1	1	221	1	1	1	714

Table 3.2: Response rates for possible outcomes

Response Rate	Exercise	PCIE	Seeking
wave 1	75.62%	99.00%	98.76%
wave 2	61.84%	63.28%	63.63%
wave 3	53.68%	55.67%	55.87%

Table 3.3: Patient characteristics by response time

Response	Wave 1 (n=1520)		Wave 2 (n=1243)		Wave 3 (n=1079)	
	%	N	%	N	%	N
Type of Cancer						
Colorectal	35.26%	536	31.86%	396	30.21%	326
Breast	33.09%	503	34.92%	434	35.68%	385
Prostate	31.63%	481	33.23%	413	34.11%	368
Gender						
Male	49.28%	749	48.83%	607	48.29%	521
Female	50.72%	771	51.17%	636	51.71%	558
Stage						
.	6.05%	92	6.11%	76	6.21%	67
0	9.14%	139	8.21%	102	10.01%	108
1	17.11%	260	18.99%	236	19.18%	207
2	38.55%	586	42.24%	525	44.49%	480
3	12.11%	184	12.15%	151	11.49%	124
4	17.04%	259	12.31%	153	8.62%	93
Age	Mean	Median (Range)	Mean	Median (Range)	Mean	Median (Range)
	64.26	65 (23–98)	63.90	64 (24–103)	63.49	64 (27–103)
PCIE Score	Mean	Median (Range)	Mean	Median (Range)	Mean	Median (Range)
	-0.006	0.0006 (-1.274–1.141)	-0.004	-0.212 (-0.702–1.886)	0.007	-0.226 (-0.604 –2.040)

Table 3.4: Longitudinal analysis of PCIE data

	Full-likelihood AR1					TPL AR1					IPL					Weighted Gee					GEE				
	Estimate	S.E	95%L	95%U		Estimate	S.E	95%L	95%U		Estimate	S.E	95%L	95%U		Estimate	S.E	95% L	95% U		Estimate	S.E	95%L	95%U	
Intercept	3.711	0.403	2.921	4.500	*	3.665	0.420	2.842	4.489	*	3.491	0.432	2.645	4.337	*	3.540	0.424	2.708	4.372	*	3.580	0.410	2.775	4.385	*
TypeCancer.Breast	0.017	0.141	-0.260	0.295		0.012	0.140	-0.262	0.286		-0.025	0.143	-0.307	0.256		0.044	0.143	-0.237	0.325		0.006	0.139	-0.266	0.278	
TypeCancer.Prostate	0.192	0.155	-0.111	0.496		0.162	0.160	-0.151	0.475		0.138	0.168	-0.192	0.468		0.190	0.165	-0.135	0.514		0.171	0.160	-0.142	0.484	
Gender	-0.282	0.154	-0.584	0.020		-0.293	0.161	-0.609	0.023		-0.303	0.163	-0.621	0.016		-0.284	0.164	-0.606	0.038		-0.269	0.159	-0.581	0.042	
Diagnosis Age	0.010	0.004	0.003	0.018	*	0.010	0.004	0.003	0.018	*	0.011	0.004	0.004	0.019	*	0.011	0.004	0.003	0.019	*	0.010	0.004	0.003	0.018	*
Stage.1	-0.103	0.211	-0.517	0.310		-0.087	0.209	-0.496	0.323		-0.154	0.223	-0.591	0.284		-0.117	0.218	-0.544	0.310		-0.103	0.213	-0.521	0.315	
Stage.2	0.032	0.218	-0.396	0.460		0.049	0.218	-0.378	0.476		0.012	0.230	-0.438	0.463		-0.006	0.230	-0.457	0.446		0.029	0.223	-0.408	0.465	
Stage.3	-0.358	0.246	-0.841	0.125		-0.344	0.245	-0.825	0.136		-0.330	0.257	-0.834	0.175		-0.357	0.248	-0.844	0.130		-0.346	0.245	-0.826	0.134	
Stage.4	-0.567	0.259	-1.074	-0.059	*	-0.549	0.260	-1.059	-0.040	*	-0.469	0.270	-0.998	0.060		-0.485	0.275	-1.024	0.053		-0.525	0.259	-1.033	-0.018	
Wave	-0.069	0.046	-0.159	0.022		-0.053	0.043	-0.137	0.032		0.142	0.109	-0.072	0.356		0.041	0.030	-0.018	0.100		0.040	0.030	-0.018	0.098	
PCIE	-0.005	0.060	-0.122	0.112		0.001	0.059	-0.115	0.116		0.160	0.068	0.027	0.293	*	-0.014	0.056	-0.124	0.096		-0.007	0.054	-0.112	0.098	
$\boldsymbol{\sigma}$	2.103	0.020	2.102	2.104		2.098	0.017	2.097	2.099		2.086	0.018	2.086	2.087											
$\boldsymbol{\rho}$	0.633	0.025	0.632	0.634		0.629	0.037	0.626	0.631																
Missing data model																									
$y_i(\vartheta_c)$	-0.142	0.041	-0.223	-0.061	*	-0.113	0.038	-0.188	-0.037	*	0.063	0.121	-0.174	0.301											
$r_{i-1}(\psi_c)$	-0.792	0.082	-0.954	-0.631	*	-0.802	0.069	-0.938	-0.666	*															
π	0.824	0.023	0.823	0.825		0.824	0.023	0.823	0.825		0.824	0.023	0.823	0.825											
Intercept	-0.472	0.419	-1.294	0.350		-0.560	0.451	-1.445	0.324		-2.141	0.669	-3.453	-0.830	*										
TypeCancer.Breast	-0.170	0.121	-0.408	0.067		-0.170	0.136	-0.436	0.095		-0.155	0.142	-0.433	0.122											
TypeCancer.Prostate	-0.301	0.150	-0.595	-0.007	*	-0.304	0.163	-0.623	0.015		-0.298	0.170	-0.632	0.036											
Gender	-0.176	0.126	-0.424	0.071		-0.169	0.143	-0.449	0.112		-0.108	0.156	-0.415	0.198											
Diagnosis Age	0.011	0.004	0.004	0.019	*	0.011	0.004	0.003	0.019	*	0.010	0.004	0.002	0.019	*										
Stage.1	-0.208	0.250	-0.698	0.281		-0.207	0.261	-0.718	0.304		-0.210	0.273	-0.745	0.326											
Stage.2	-0.163	0.220	-0.594	0.269		-0.164	0.231	-0.617	0.288		-0.199	0.235	-0.659	0.262											
Stage.3	-0.048	0.250	-0.539	0.443		-0.036	0.263	-0.550	0.479		-0.006	0.278	-0.552	0.539											
Stage.4	0.710	0.264	0.192	1.227	*	0.726	0.273	0.191	1.261	*	0.783	0.311	0.173	1.392	*										
Wave	0.280	0.073	0.137	0.423	*	0.283	0.062	0.161	0.405	*	0.393	0.051	0.292	0.494	*										
PCIE	-0.026	0.260	-0.536	0.483		-0.033	0.258	-0.539	0.474		-0.066	0.271	-0.596	0.465											

[1] σ standard deviation of exercise at each wave. [2] ρ pairwise-correlation of exercise at each adjacent wave. [3] TPL: transition pseudo-likelihood
[4] IPL: Independent pseudo-likelihood

Table 3.5: Simulation study of sensitivity to sample size, 1000 replicates

N=300	Toep(2)		TPL			IPL		
Resp Rate								
0.911	0.747	0.650						
		True	Bias	Efficiency	C.P	Bias	Efficiency	C.P
	α_0	11.500	0.001	0.974	0.948	0.002	1.000	0.949
	α_1	1.050	0.014	0.984	0.948	0.001	1.000	0.952
	α_2	0.250	0.004	0.974	0.953	0.004	1.000	0.952
	α_3	0.210	0.001	0.979	0.946	0.001	1.000	0.947
	β_0	-14.500	0.017	1.080	0.940	1.876	1.000	0.856
	β_1	-0.150	0.004	1.258	0.944	0.009	1.000	0.805
	β_2	-0.550	0.056	0.940	0.951	0.082	1.000	0.948
	β_3	-0.200	0.044	0.897	0.945	0.337	1.000	0.927
	β_4	0.310	0.030	0.903	0.953	0.028	1.000	0.949
	$\beta_5(y_t)$	1.450	0.008	0.941	0.938	0.033	1.000	0.918
	$\beta_6(r_{t-1})$	1.200	0.101		0.945			
	σ	1.200	0.006	0.968	0.940	0.005	1.000	0.945
	ρ	0.400	0.003		0.955			
	π	0.910	0.001	1.000	0.957	0.001	1.000	0.957
N=500	Toep(2)							
Resp Rate								
0.910	0.746	0.650						
		True	Bias	Efficiency	C.P	Bias	Efficiency	C.P
	α_0	11.500	0.004	0.973	0.934	0.002	1.000	0.930
	α_1	1.050	0.016	0.984	0.932	0.000	1.000	0.946
	α_2	0.250	0.003	0.974	0.937	0.002	1.000	0.936
	α_3	0.210	0.000	0.978	0.931	0.000	1.000	0.931
	β_0	-14.500	0.187	1.080	0.942	2.059	1.000	0.834
	β_1	-0.150	0.001	1.264	0.951	0.012	1.000	0.779
	β_2	-0.550	0.054	0.941	0.948	0.006	1.000	0.948
	β_3	-0.200	0.049	0.899	0.955	0.329	1.000	0.898
	β_4	0.310	0.012	0.903	0.946	0.005	1.000	0.955
	$\beta_5(y_t)$	1.450	0.013	0.942	0.932	0.055	1.000	0.923
	$\beta_6(r_{t-1})$	1.200	0.102		0.930			
	σ	1.200	0.005	0.967	0.937	0.003	1.000	0.942
	ρ	0.400	0.002		0.932			
	π	0.910	0.000	1.000	0.956	0.000	1.000	0.956
N=1000	Toep(2)							
Resp Rate								
0.910	0.747	0.651						
		True	Bias	Efficiency	C.P	Bias	Efficiency	C.P
	α_0	11.500	0.003	0.973	0.944	0.001	1.000	0.948
	α_1	1.050	0.015	0.984	0.928	0.000	1.000	0.944
	α_2	0.250	0.001	0.974	0.953	0.000	1.000	0.951
	α_3	0.210	0.001	0.979	0.949	0.000	1.000	0.955
	β_0	-14.500	0.234	1.076	0.948	2.046	1.000	0.773
	β_1	-0.150	0.000	1.258	0.941	0.013	1.000	0.660
	β_2	-0.550	0.058	0.937	0.948	0.090	1.000	0.947
	β_3	-0.200	0.029	0.896	0.947	0.348	1.000	0.815
	β_4	0.310	0.021	0.899	0.949	0.023	1.000	0.944
	$\beta_5(y_t)$	1.450	0.015	0.937	0.941	0.053	1.000	0.937
	$\beta_6(r_{t-1})$	1.200	0.110		0.912			
	σ	1.200	0.003	0.965	0.950	0.002	1.000	0.946
	ρ	0.400	0.001		0.955			
	π	0.910	0.000	1.000	0.959	0.000	1.000	0.959

[1] Resp Rate: response rate; C.P.: coverage probability.
[2] σ standard deviation of outcome at each time.
[3] ρ pairwise-correlation of outcome at each adjacent pairs.

Table 3.6: Simulation study of model comparison, 500 replicates

	N=300													
Resp Rate time 1		0.910							0.910					
Resp Rate time 2		0.747							0.817					
Resp Rate time 3		0.651							0.705					
		TPL Model Correctly specified			IPL model				TPL Model				IPL model Correctly specified	
	TRUE	E.est	S.D	C.P	E.est	S.D	C.P	TRUE	E.est	S.D	C.P	E.est	S.D	C.P
α_0	11.500	11.502	0.326	0.930	11.500	0.328	0.930	11.500	11.493	0.327	0.944	11.493	0.329	0.938
α_1	1.050	1.067	0.286	0.928	1.052	0.287	0.936	1.050	1.056	0.283	0.934	1.047	0.284	0.930
α_2	0.250	0.245	0.379	0.930	0.247	0.381	0.926	0.250	0.257	0.379	0.952	0.255	0.381	0.956
α_3	0.210	0.211	0.331	0.960	0.211	0.333	0.954	0.210	0.207	0.328	0.956	0.210	0.329	0.950
β_0	-14.500	-14.728	1.911	0.946	-12.900	1.867	0.894	-14.500	-14.525	1.976	0.942	-15.379	2.047	0.966
β_1	-0.150	-0.156	0.141	0.950	-0.143	0.133	0.842	-0.150	-0.155	0.146	0.946	-0.158	0.150	0.956
β_2	-0.550	-0.817	1.328	0.954	-0.889	1.350	0.950	-0.550	-0.552	1.364	0.950	-0.544	1.381	0.952
β_3	-0.200	-0.238	0.754	0.956	-0.638	0.775	0.920	-0.200	-0.082	0.767	0.930	-0.158	0.777	0.932
β_4	0.310	0.420	0.827	0.950	0.437	0.849	0.950	0.310	0.327	0.851	0.950	0.315	0.860	0.956
$\beta_5(y_t)$	1.450	1.496	0.561	0.950	1.461	0.567	0.936	1.450	1.453	0.577	0.938	1.525	0.609	0.954
$\beta_6(r_{t-1})$	1.200	1.102	0.711	0.916				0.000	-0.162	0.715	0.928			
σ	1.200	1.195	0.187	0.944	1.197	0.189	0.942	1.200	1.191	0.185	0.936	1.194	0.187	0.944
ρ	0.400	0.396	0.198	0.944				0.400	0.396	0.194	0.942			
π	0.910	0.910	0.128	0.936	0.910	0.128	0.936	0.910	0.910	0.128	0.936	0.910	0.128	0.936

[1] Simulation sample size N=300; Resp Rate: response rate.
[2] E.est:Mean estimation; S.D: Standard deviation; C.P.: coverage probability.
[3] σ standard deviation of outcome at each time. [4] ρ pairwise-correlation of outcome at each adjacent pairs.

Table 3.7: Simulation study of sensitivity to different covariance structures, 1000 replicates

			TPL			IPL		
AR1 Resp Rate 0.911	0.747	0.650						
		True	Bias	Efficiency	Coverage Prob	Bias	Efficiency	Coverage Prob
	α_0	11.500	0.001	0.983	0.944	0.002	1.000	0.944
	α_1	1.050	0.014	0.977	0.951	0.001	1.000	0.947
	α_2	0.250	0.004	0.983	0.952	0.003	1.000	0.942
	α_3	0.210	0.000	0.972	0.945	0.001	1.000	0.947
	β_0	-14.500	0.034	1.068	0.945	1.866	1.000	0.859
	β_1	-0.150	0.004	1.252	0.948	0.009	1.000	0.824
	β_2	-0.550	0.051	0.941	0.945	0.074	1.000	0.936
	β_3	-0.200	0.049	0.896	0.942	0.336	1.000	0.932
	β_4	0.310	0.028	0.904	0.952	0.025	1.000	0.950
	$\beta_5(y_t)$	1.450	0.006	0.932	0.935	0.033	1.000	0.927
	$\beta_6(r_{t-1})$	1.200	0.098		0.935			
	σ	1.200	0.006	0.966	0.935	0.005	1.000	0.941
	ρ	0.400	0.003		0.952			
	π	0.910	0.001	1.000	0.957	0.001	1.000	0.957
TOEP(2) Resp Rate 0.911	0.747	0.650						
		True	Bias	Efficiency	Coverage Prob	Bias	Efficiency	Coverage Prob
	α_0	11.500	0.001	0.974	0.948	0.002	1.000	0.949
	α_1	1.050	0.014	0.984	0.948	0.001	1.000	0.952
	α_2	0.250	0.004	0.974	0.953	0.004	1.000	0.952
	α_3	0.210	0.001	0.979	0.946	0.001	1.000	0.947
	β_0	-14.500	0.017	1.080	0.940	1.876	1.000	0.856
	β_1	-0.150	0.004	1.258	0.944	0.009	1.000	0.805
	β_2	-0.550	0.056	0.940	0.951	0.082	1.000	0.948
	β_3	-0.200	0.044	0.897	0.945	0.337	1.000	0.927
	β_4	0.310	0.030	0.903	0.953	0.028	1.000	0.949
	$\beta_5(y_t)$	1.450	0.008	0.941	0.938	0.033	1.000	0.918
	$\beta_6(r_{t-1})$	1.200	0.101		0.945			
	σ	1.200	0.006	0.968	0.940	0.005	1.000	0.945
	ρ	0.400	0.003		0.955			
	π	0.910	0.001	1.000	0.957	0.001	1.000	0.957
EXCH Resp Rate 0.911	0.748	0.650						
		True	Bias	Efficiency	Coverage Prob	Bias	Efficiency	Coverage Prob
	α_0	11.500	0.001	0.999	0.946	0.002	1.000	0.950
	α_1	1.050	0.015	0.966	0.933	0.001	1.000	0.940
	α_2	0.250	0.001	0.998	0.942	0.001	1.000	0.941
	α_3	0.210	0.003	0.961	0.943	0.003	1.000	0.943
	β_0	-14.500	0.032	1.052	0.943	1.902	1.000	0.866
	β_1	-0.150	0.003	1.246	0.941	0.010	1.000	0.806
	β_2	-0.550	0.073	0.950	0.951	0.094	1.000	0.946
	β_3	-0.200	0.052	0.906	0.944	0.324	1.000	0.918
	β_4	0.310	0.038	0.914	0.952	0.035	1.000	0.952
	$\beta_5(y_t)$	1.450	0.008	0.921	0.940	0.040	1.000	0.930
	$\beta_6(r_{t-1})$	1.200	0.099		0.928			
	σ	1.200	0.007	0.967	0.943	0.005	1.000	0.943
	ρ	0.400	0.006		0.962			
	π	0.910	0.001	1.000	0.957	0.001	1.000	0.957

[1] Simulation sample size N=500. Resp Rate: response rate.
[2] σ standard deviation of outcome at each time. [3] ρ pairwise-correlation of outcome at each adjacent pairs.

Table 3.8: Simulation study of normal data vs gamma data, 500 replicates

$\rho = 0$

Resp.Rate	0.91	0.75	0.67											
		Normal						Gamma						
		TPL			IPL			TPL				IPL		
	TRUE	E.est	STD	Cov.Prob	E.est	STD	Cov.Prob	TRUE	E.est	STD	Cov.Prob	E.est	STD	Cov.Prob
α_0	11.500	11.502	0.346	0.964	11.503	0.346	0.962	11.500	11.494	0.346	0.946	11.495	0.346	0.944
α_1	1.050	1.019	0.323	0.934	1.017	0.322	0.940	1.050	1.002	0.316	0.922	0.999	0.315	0.918
β_0	-14.500	-15.048	1.911	0.936	-12.960	1.787	0.858	-14.500	-15.024	1.880	0.928	-12.856	1.751	0.832
β_1	-0.150	-0.156	0.164	0.944	-0.143	0.153	0.870	-0.150	-0.149	0.162	0.894	-0.136	0.150	0.794
β_2	-0.550	-0.593	1.445	0.954	-0.569	1.441	0.958	-0.550	-0.552	1.421	0.942	-0.541	1.417	0.952
β_3	-0.200	-0.126	0.795	0.944	-0.457	0.791	0.934	-0.200	-0.173	0.784	0.952	-0.479	0.779	0.936
β_4	0.310	0.317	0.900	0.956	0.305	0.903	0.954	0.310	0.306	0.886	0.942	0.298	0.889	0.954
$\beta_5(y_t)$	1.450	1.497	0.557	0.930	1.436	0.536	0.924	1.450	1.487	0.547	0.932	1.415	0.525	0.904
$\beta_6(r_{t-1})$	1.200	1.255	0.762	0.946				1.200	1.217	0.751	0.950			
π	0.910	0.910	0.128	0.946	0.910	0.128	0.946	0.910	0.911	0.128	0.938	0.911	0.128	0.938
σ_1	2.145	2.214	0.253	0.792	2.214	0.253	0.790	2.145	2.156	0.252	0.950	2.155	0.252	0.948
σ_2	2.241							2.241						
σ_3	2.332							2.332						
ρ	0.000	0.000	0.222	0.950				0.000	-0.004	0.222	0.946			

$\rho = 0.5$

Resp.Rate	0.91	0.75	0.67											
		Normal						Gamma						
		TPL			IPL			TPL				IPL		
	TRUE	E.est	STD	Cov.Prob	E.est	STD	Cov.Prob	TRUE	E.est	STD	Cov.Prob	E.est	STD	Cov.Prob
α_0	11.500	11.527	0.353	0.944	11.503	0.353	0.948	11.500	11.514	0.350	0.948	11.489	0.352	0.964
α_1	1.050	1.089	0.287	0.918	1.024	0.289	0.924	1.050	1.046	0.273	0.948	0.995	0.276	0.892
β_0	-14.500	-14.135	1.855	0.920	-12.214	1.753	0.816	-14.500	-13.141	1.768	0.850	-11.898	1.696	0.756
β_1	-0.150	-0.150	0.157	0.904	-0.140	0.150	0.828	-0.150	-0.134	0.147	0.754	-0.130	0.143	0.706
β_2	-0.550	-0.590	1.401	0.954	-0.614	1.414	0.950	-0.550	-0.436	1.350	0.972	-0.494	1.391	0.972
β_3	-0.200	-0.139	0.770	0.936	-0.487	0.783	0.946	-0.200	-0.090	0.739	0.952	-0.464	0.766	0.936
β_4	0.310	0.328	0.875	0.960	0.334	0.888	0.956	0.310	0.246	0.841	0.966	0.269	0.872	0.968
$\beta_5(y_t)$	1.450	1.437	0.533	0.906	1.377	0.528	0.886	1.450	1.314	0.501	0.814	1.327	0.508	0.864
$\beta_6(r_{t-1})$	1.200	0.878	0.768	0.894				1.200	0.913	0.736	0.884			
π	0.910	0.910	0.128	0.932	0.910	0.128	0.932	0.910	0.909	0.128	0.934	0.909	0.128	0.934
σ_1	2.145	2.213	0.271	0.854	2.214	0.274	0.864	2.145	2.127	0.268	0.932	2.142	0.272	0.940
σ_2	2.241							2.241						
σ_3	2.332							2.332						
ρ	0.500	0.496	0.204	0.942				0.500	0.522	0.212	0.884			

$\rho = 0.707$

Resp.Rate	0.91	0.75	0.67											
		Normal						Gamma						
		TPL			IPL			TPL				IPL		
	TRUE	E.est	STD	Cov.Prob	E.est	STD	Cov.Prob	TRUE	E.est	STD	Cov.Prob	E.est	STD	Cov.Prob
α_0	11.500	11.580	0.357	0.928	11.503	0.356	0.950	11.500	11.570	0.352	0.916	11.493	0.354	0.942
α_1	1.050	1.083	0.258	0.916	1.022	0.272	0.928	1.050	1.039	0.246	0.944	1.001	0.259	0.866
β_0	-14.500	-13.254	1.748	0.886	-11.753	1.745	0.752	-14.500	-12.339	1.669	0.788	-11.926	1.722	0.756
β_1	-0.150	-0.143	0.146	0.858	-0.137	0.148	0.778	-0.150	-0.127	0.138	0.652	-0.131	0.146	0.698
β_2	-0.550	-0.647	1.367	0.950	-0.695	1.402	0.952	-0.550	-0.506	1.316	0.952	-0.526	1.389	0.954
β_3	-0.200	-0.080	0.748	0.952	-0.460	0.776	0.948	-0.200	-0.104	0.718	0.922	-0.479	0.769	0.922
β_4	0.310	0.333	0.855	0.958	0.351	0.880	0.958	0.310	0.286	0.821	0.944	0.302	0.872	0.948
$\beta_5(y_t)$	1.450	1.355	0.492	0.836	1.331	0.526	0.858	1.450	1.251	0.466	0.726	1.331	0.518	0.860
$\beta_6(r_{t-1})$	1.200	0.714	0.762	0.846				1.200	0.708	0.728	0.810			
π	0.910	0.909	0.128	0.958	0.909	0.128	0.958	0.910	0.910	0.128	0.930	0.910	0.128	0.930
σ_1	2.145	2.213	0.283	0.876	2.218	0.290	0.874	2.145	2.119	0.275	0.924	2.146	0.282	0.948
σ_2	2.241													
σ_3	2.332													
ρ	0.707	0.704	0.169	0.946				0.707	0.711	0.183	0.934			

[1] Simulation sample size N=300. [2] σ_j standard deviation of outcome at wave j. [3] ρ pairwise-correlation of outcome at each adjacent pairs.

CHAPTER 4 : A hidden Markov model for non-ignorable non-monotone missing longitudinal data for medical studies of quality of life

4.1. Introduction

In a longitudinal study, subjects are observed as time progresses. A common problem is that repeated measurements are not fully observed due to missing responses or loss to follow up. Individuals can move in and out of the observed data set, giving rise to a large class of distinct "non-monotone" missingness patterns. The appropriate statistical methods differ according to the data structure and missingness mechanism. When the missingness is MCAR (missing completely at random) or MAR (missing at random), data analysis is the most straightforward. Little and Rubin (1987) and Allison (2001) provide helpful terminology to describe missing data mechanisms and a comprehensive overview of potential methods. Most approaches can be categorized as selection models, pattern-mixture models or shared-parameter models depending on the factorization of the joint likelihood of the outcomes and missingness indicators. Multi-state Markov models, on the other hand, are commonly used to describe disease progression studies (Commenges et al., 2004; Jackson et al., 2003), and observational studies in cancer (Sutradhar et al., 2010; Uhry et al., 2010). Wall and Li (2009) and Cooper and Lipsitch (2004) extended multi-state Markov models to hidden Markov models to obtain a more flexible transition matrix. Maruotti (2011) and Altman (2007) provided a good review of methodology for use of hidden Markov models in the longitudinal data framework.

In chronic disease studies, longitudinal data can be used to monitor disease progression. In health care survey studies, longitudinal data can be used to measure changes in attitude or compliance with treatment or medical advice. The underlying structure of longitudinal data can be complicated due to the fact that during follow-up, the occurrence of observations at a given time depend on unobserved (hidden) states such as changes in disease condition, recovery, progression, or better access to health care. Thus both repeated assess-

ments and missingness could depend on the current hidden state. A common assumption in these studies is that the missing assessment data at each time are non-informative. If true, modeling observed data directly with assumption of MCAR data will provide unbiased estimation. Scott et al. (2005) developed a hidden Markov model for medical longitudinal data using k-means clustering analysis in a traditional health state model assuming an ignorable missingness mechanism. However, under a longitudinal scheme, observations are recorded at periodic times, depending on hidden states which often do have a well defined meaning at given time. The missingness mechanism may depend on recorded assessments, the hidden states, or a combination of them. Modeling such data without considering the missingness will result in a biased estimation (Ibrahim and Molenberghs, 2009; Troxel et al., 1998b). Many proposed methods have been developed to deal with monotone missingness patterns (Spagnoli et al., 2011; Ie Cessie et al., 2009; Philipson et al., 2008), by incorporating the missingness indicator into the transition matrix. However, there is little work that addresses "non-monotone" and "non-ignorable" missingness in Markov process models. Sweeting et al. (2009) presented a partially hidden Markov model using observed auxiliary variables to model "non-monotone" and "non-ignorable" missingness patterns for disease progression. This model is inefficient, however, if the correlation between the auxiliary variables and outcome becomes weak; often such auxiliary variables do not exist, and the assumption itself is hard to examine. Chen et al. (2010) proposed a piecewise constant transition model to address multi-state Markov model assuming non-homogeneous Markov process. Their primary interest is in continuous-time multi-state model parameters and transition intensities. Chen and Zhou (2011) extended the work to non-parametric time-transformation models to make the model more flexible.

We propose a method assuming a time-homogeneous hidden Markov process and mainly focus on discrete hidden states. We treat the initial probability and transition matrix as nuisance parameters since the primary interest is in parameters in the state-dependent model and the missingness mechanism model. The proposed two-stage pseudo-likelihood method (Gong and Samaniego, 1981; Parke, 1986) updates the nuisance parameter using

"convenient" estimation via the *backward-forward* algorithm (Baum et al., 1970; Rabiner, 1989; Welch, 2003). By employing the quasi-Newton algorithm, we maximize the pseudo-likelihood function to update the estimation iteratively. The Levenberg-Marquardt algorithm (Turner, 2008), a modified Newton's method, is used to achieve better parameter estimation accuracy. Sandwich estimators are used to recover robust covariance estimation (Liang and Zeger, 1986) and confidence intervals. The AIC/BIC criterion could be used to defined the "best" number of hidden states. However, caution is needed since this method has not been justified theoretically in this context. MacKay (2002) gives a discussion and an alternative model selection criterion in the simple hidden Markov model. Comparing with other methods, our proposed method has no need to pre-specify the underlying transition matrix. Guihenneuc-Jouyaux et al. (2000), and Sabin et al. (1996) showed the estimation in the hidden Markov model can be inefficient if the pre-specified transition matrix departs from the true underlying transition matrix. Secondly, our proposed model does not increase the parameter space as fast as other methods when the number of hidden states increases, which makes the model estimation more appealing.

In this paper, we will introduce a recent application in Section 4.2, describe the proposed methods in Section 4.3, present a simulation study to address the performance of the methods in Section 4.4, and summarize our analysis of the data set in Section 4.5. Section 4.6 provides a discussion, some brief comments and ideas for future work.

4.2. Motivating Example

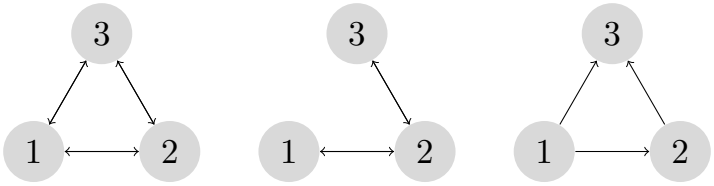
We consider data from a non-blinded randomized phase III intergroup trial (RTOG 9402) evaluating the overall survival of patients with anaplastic oligodendroglioma (AO) brain tumors who received either chemotherapy plus radiation therapy (Arm 1) or radiation therapy alone (Arm 2), previously described by Cairncross et al. (2006) and Wang et al. (2010). Studies show that AO patients respond to surgery and radiotherapy (RT) at diagnosis, as well to procarbazine, lomustine, and vincristine (PCV) chemotherapy; it was unclear whether patients would benefit from combined PCV and RT therapy, compared to RT only.

Coordinated by the National Cancer Institute, the Radiation Therapy Oncology Group (RTOG) conducted this randomized trial (9402) between 1994 and 2002. The study enrolled 289 eligible participants. Study reports showed that patients who have the 1p and 19q deletions receive significantly longer progression survival times regardless treatment, but this is associated with substantial toxicity. There was no significant difference in median survival times between two treatment arms in patients with only one deletion or no deletions of chromosomal segment.

The effect of toxicity and side effects from PCV chemotherapy and RT on patients' neurologic functioning and global quality of life remain unclear. Several measures were collected at each visit to assess patients' cognitive ability and attitude on quality of life during the studying time period, including Karnofsky performance status (KPS), which measures physical well-being; the Mini-Mental status exam (MMSE), which measures cognitive ability as assessed by a nurse, research associate, or physician to reflect the opinions of health care specialist; and the modified Brain Quality of Life Questionnaire (B-QLQ), which measures patient-reported quality of life. If a patient required help to finish the B-QLQ questionnaire, the reasons were documented.

It makes sense that patients' functional status may depend on their underlying health status. For example, patients may feel better after their cancer responds to treatment. The outcome process may be influenced by both known and unknown medical/non-medical information (hidden states). A first-order hidden Markov dependence structure fits the natural data structure in the longitudinal framework. For example, let $S = 3$ be the number of hidden states, with 1 = **stable**, 2 = **relapse**, 3 = **crisis**; the MMSE and B-QLQ scores, can depend on the actual states at a given time. The missingness mechanism is conditional on a function of both assessments and unobserved hidden states S. The diagram below indicates the possible relationships among different underlying hidden transition states associated

with assessments.



Transition matrix $\mathbb{Q}$ for $j, k \in (1, 2, 3)$

$$\mathbb{Q}_1 = \left\{ \begin{array}{ccc} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{array} \right\}_{P6} \quad \mathbb{Q}_2 = \left\{ \begin{array}{ccc} q_{11} & q_{12} & 0 \\ q_{21} & q_{22} & q_{23} \\ 0 & q_{32} & q_{33} \end{array} \right\}_{P4} \quad \mathbb{Q}_3 = \left\{ \begin{array}{ccc} q_{11} & q_{12} & q_{13} \\ 0 & q_{22} & q_{23} \\ 0 & 0 & 1 \end{array} \right\}_{P3}$$

with

$$\sum_{k=1}^{S} q_{jk} = 1, \quad S \in I_s \quad I_s = (1, 2, \ldots, S)$$

$\mathbb{Q}_1$ is an unconstrained, fully connected or ergodic transition matrix in which transitions are possible between any two states. $\mathbb{Q}_2$ is a first-order symmetric transition matrix, in which transitions only occur between adjacent states. $\mathbb{Q}_3$ could be described as an illness to death model, in which patients progress to the next state but never recover. Clearly the estimation of Markov chain parameters becomes more complicated as the number of hidden states increases. However, by modeling the disease progression through different hidden states, our approach allows more precise identification of the treatment effect. On the other hand, too many hidden states make the application difficult to estimate and interpret.

4.3. Methods and Notation

4.3.1. Notation and underlying assumptions

Instead of observing all Y_{it}, with $t \in (1 \cdots T)$ assessment times, we observe the below pseudo-observations $O_i = (Y_i, R_i)$

$$\begin{array}{ccccccccc} & O_{i1} & \cdots & O_{i2} & \cdots & O_{i,T-1} & \cdots & O_{iT} & . \\ & \uparrow & & \uparrow & & \uparrow & & \uparrow & \\ \xrightarrow{\pi_{ij}} & S_{i1} & \xrightarrow{q_{itjk}} & S_{i2} & \cdots & S_{i,T-1} & \xrightarrow{q_{itjk}} & S_{iT} & \end{array}$$

The pseudo-observations $O_i = (Y_i, R_i)$ are conditionally independent, i.e., $f(O_{is}|S_{is} = j) \perp f(O_{it}|S_{it} = k)$; as above $i \in (1 \cdots N)$ denotes subjects, and $j, k \in (1 \cdots S)$ denote hidden states. Each element of the missingness indicator vector R_i is defined as

$$R_{it} = \begin{cases} 0 & \text{if} \quad \text{missing } Y_{it} \\ 1 & \text{if} \quad \text{observed } Y_{it} \end{cases} \quad .$$

The simplest model in this framework is the homogeneous hidden Markov model, which assumes a stationary Markov transition probability $q_{itjk} = q_{jk}$ and a common initial probability $\pi_{ij} = \pi_j$, where $i \in (1, 2, \cdots N)$ denotes subjects, and $j, k \in (1, 2, \cdots, S)$ denote hidden states. A simple two state transition matrix $\mathbb{Q}$ for $j, k \in (1, 2)$ is

$$\mathbb{Q} = \begin{Bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{Bmatrix} \quad ,$$

with defined transition probability $q_{jk} = f(s_{i,t+1} = k|s_{i,t} = j) = f(s_{i,t+2} = k|s_{i,t+1} = j)$ and initial probability $\pi_j = f(s_{i1} = j)$; these satisfy the conditions $\sum_{k=1}^{S} q_{jk} = 1$ and $\sum_{j=1}^{S} \pi_j = 1$, where $S \in I_s$, $I_s = (1, 2, \ldots, S)$.

The conditional density $f(O_{it}|S_{it} = j)$ follows an independent Bernoulli distribution with

density function

$$f(O_{it}|S_{it}=j)=\begin{cases}\int_{Y_{it}} f(Y_{it},R_{it}|S_{it})dy_{it} & \text{if} \quad R_{it}=0\\ f(Y_{it},R_{it}|S_{it}) & \text{if} \quad R_{it}=1\end{cases} .$$

4.3.2. Selection hidden Markov model (SHMM)

Selection models (Little and Rubin, 1987; Allison, 2001) are a commonly used approach to non-ignorable missingness in longitudinal data. Selection models can be written as the joint distribution of Y_i and R_i in the form

$$\begin{aligned} f(Y_{i,obs},R_i|X_i,\vartheta) &= \int f(Y_{i,obs},Y_{i,mis},R_i|X_i,\vartheta)dY_{i,mis} \\ &= \int f(Y_{i,obs},Y_{i,mis}|X_i,\alpha)f(R_i|Y_{i,obs},Y_{i,mis},X_i,\beta)dY_{i,mis}\end{aligned}$$

where $Y_i=(y_{i1},y_{i2},\cdots y_{it})$, $R_i=(r_{i1},r_{i2},\cdots r_{it})$, $\vartheta=(\alpha,\beta)$. A selection model can be easily combined with a hidden Markov model as described in the next section.

Outcomes dependent missingness

In this scenario, the missingness of an observation depends only on outcomes. We define the conditional density of $Y_{it},R_{it}|S_{it}$ as

$$f(Y_{it},R_{it}|S_{it}=j)=f(Y_{it}|S_{it}=j)*f(R_{it}|Y_{it}).$$

The conditional observation $\{Y_{it}|S_{it}=j\}$ is i.i.d. from an exponential family where

$$\begin{aligned} f(y_{it}|s_{it}=j,\alpha) &= \exp\{(y_{it}\eta_{itj}-c(\eta_{itj}))/a(\phi)+d(y_{iy},\phi)\} \\ \eta_{itj} &= \alpha_{j0}+\boldsymbol{\alpha}_j^{'}\mathbf{x}_{it}\end{aligned}$$

with the missingness indicator $\{R_{it}|Y_{it}\}$ following a Bernoulli distribution modeled as

$$logit(Pr(R_{it} = 1|Y_{it})) = \beta_0 + \boldsymbol{\beta}_1^{'}\mathbf{x}_{it} + \beta_2 * Y_{it}.$$

Here $\mathbf{x}_{it}$ is a time dependent covariate matrix, and $\boldsymbol{\alpha}_j, \boldsymbol{\beta}_1$ are the corresponding parameter vectors. Testing $\beta_2 \neq 0$ is equivalent to checking if the missing data are non-ignorable.

State dependent missingness

In this scenario, the missingness of an observation depends on a function of outcome and hidden states. Define the conditional density of $Y_{it}, R_{it}|S_{it}$ as

$$f(Y_{it}, R_{it}|S_{it} = j) = f(Y_{it}|S_{it} = j) * f(R_{it}|Y_{it}, S_{it} = j)$$

As above the conditional observation $\{Y_{it}|S_{it} = j\}$ is i.i.d. from an exponential family where

$$\begin{aligned} f(y_{it}|s_{it} = j, \alpha) &= \exp\{(y_{it}\eta_{itj} - c(\eta_{itj}))/a(\phi) + d(y_{iy}, \phi)\} \\ \eta_{itj} &= \alpha_{j0} + \boldsymbol{\alpha}_j^{'}\mathbf{x}_{it} \end{aligned}$$

with the missingness indicator $\{R_{it}|Y_{it}, S_{it} = j\}$ following a Bernoulli distribution modeled as

$$logit(Pr(R_{it} = 1|Y_{it}, S_{it} = j)) = \beta_{j0} + \boldsymbol{\beta}_{j1}^{'}\mathbf{x}_{it} + \beta_{j2} * Y_{it}.$$

Clearly the parameters $\beta_0, \boldsymbol{\beta_1}, \beta_2$ are the average effects of the parameters $\beta_{j0}, \boldsymbol{\beta_{j1}}, \beta_{j2}$. We can test each $\beta_{j2} \neq 0, \quad j \in (1, 2, \cdots S)$ to check if the missing data are non-ignorable. In practice, outcome dependent missingness models are likely more useful since the primary interest here are the state-dependent model coefficients, and there are fewer parameters to be estimated in the marginal model.

4.3.3. Shared parameter hidden Markov model (SPHMM)

Shared parameter models (Gao, 2004; Alfo and Maruotti, 2009) form another class of approaches to dealing with non-ignorable missing data by introducing a shared latent quantity to factorize the joint density, as follows:

$$\begin{aligned} f(Y_{i,obs}, R_i|\vartheta) &= \int f(Y_{i,obs}, Y_{i,mis}, R_i|\vartheta) dY_{i,mis} \\ &= \int\int f(Y_{i,obs}, Y_{i,mis}, R_i|b_i, \alpha, \beta) * f(b_i|\psi) db_i dY_{i,mis} \\ &= \int\int f(Y_{i,obs}, Y_{i,mis}|b_i, \alpha) dY_{i,mis} f(R_i|b_i, \beta) * f(b_i|\psi) db_i \\ &= \int f(Y_{i,obs}|b_i, \alpha) f(R_i|b_i, \beta) * f(b_i|\psi) db_i \end{aligned}$$

where $Y_i = (y_{i1}, y_{i2}, \ldots y_{it})$, $R_i = (r_{i1}, r_{i2}, \ldots r_{it})$. Shared parameter models assume independence between the outcome process and the missing indicators conditional on the shared parameter b_i. Similarly, shared parameter models can work with hidden Markov models easily.

Let the conditional density $f(O_{it}|S_{it}, b_i)$ follow a Bernoulli distribution denoted as

$$f(O_{it}|S_{it} = j, b_i) = \begin{cases} \int_{Y_{it}} f(Y_{it}, R_{it}|S_{it}, b_i) dy_{it} & \text{if} \quad R_{it} = 0 \\ f(Y_{it}, R_{it}|S_{it}, b_i) & \text{if} \quad R_{it} = 1 \end{cases}$$

which can be simplified as

$$f(O_{it}|S_{it} = j, b_i) = \begin{cases} f(R_{it}|S_{it}, b_i) & \text{if} \quad R_{it} = 0 \\ f(Y_{it}|S_{it}, b_i) * f(R_{it}|S_{it}, b_i) & \text{if} \quad R_{it} = 1 \end{cases} .$$

To further simplify the model, we assume the shared latent variables (random effects b_i) are independent with hidden states

$$f(S_{it} = s_{it}|b_i) = f(S_{it} = s_{it}) \quad .$$

Similarly, as in the SHMM, we can define the main model and missingness mechanism model described in the next section.

Outcome dependent missingness

In this scenario, the missingness of observations depends on outcomes and the random effect. We define the conditional density of $Y_{it}, R_{it}|S_{it}, b_i$ as

$$f(Y_{it}, R_{it}|S_{it} = j, b_i) = f(Y_{it}|S_{it} = j, b_i) * f(R_{it}|b_i) \quad .$$

The conditional observation $\{Y_{it}|S_{it} = j, b_i\}$ is i.i.d. from an exponential family where

$$\begin{aligned} f(y_{it}|s_{it} = j, \alpha, b_i) &= \exp\{(y_{it}\eta_{itj} - c(\eta_{itj}))/a(\phi) + d(y_{iy}, \phi)\} \\ \eta_{itj} &= \alpha_{j0} + \boldsymbol{\alpha}_j^{'}\mathbf{x}_{it} + \mathbf{b_i}^{'}\mathbf{z}_{it} \end{aligned}$$

with the missingness indicators $\{R_{it}|Y_{it}, b_i\}$ following a Bernoulli distribution modeled as

$$logit(Pr(R_{it} = 1|Y_{it}, b_i)) = \beta_0 + \boldsymbol{\beta}_1^{'}\mathbf{x}_{it} + \mathbf{b_i}^{'}\mathbf{z}_{it} \quad .$$

Here $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ are time-dependent covariate matrices for fixed and random effects; $\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \mathbf{b}_i$ are the corresponding parameter vectors. Testing $\mathbf{b_i} \neq 0$ is equivalent to checking if the missing data are non-ignorable.

State dependent missingness

In this scenario, the missingness of observations depends on a function of outcomes, hidden states and the random effects. We define the conditional density of $Y_{it}, R_{it}|S_{it}, b_i$ as

$$f(Y_{it}, R_{it}|S_{it} = j, b_i) = f(Y_{it}|S_{it} = j, b_i) * f(R_{it}|S_{it} = j, b_i) \quad .$$

The conditional observation $\{Y_{it}|S_{it} = j, b_i\}$ is i.i.d. from an exponential family where

$$\begin{aligned} f(y_{it}|s_{it} = j, \alpha, b_i) &= \exp\{(y_{it}\eta_{itj} - c(\eta_{itj}))/a(\phi) + d(y_{iy}, \phi)\} \\ \eta_{itj} &= \alpha_{j0} + \boldsymbol{\alpha}_j^{'}\mathbf{x}_{it} + \mathbf{b_i}^{'}\mathbf{z}_{it} \end{aligned}$$

with the missingness indicators $\{R_{it}|Y_{it}, s_{it} = j, \mathbf{b_i}\}$ following a Bernoulli distribution

$$logit(Pr(R_{it} = 1|Y_{it}, s_{it} = j, \mathbf{b_i})) = \beta_{j0} + \boldsymbol{\beta}_{j1}^{'}\mathbf{x}_{it} + \mathbf{b_i}^{'}\mathbf{z}_{it} \quad .$$

Clearly the parameters β_0 and $\boldsymbol{\beta}_1$ are the average effects of the parameters β_{j0} and $\boldsymbol{\beta}_{j1}$. Testing $\mathbf{b_i} \neq 0$ is equivalent to checking if the missing data are non-ignorable. In practice, the random effect can be treated as nuisance parameter, like the transition matrix and initial probabilities.

4.3.4. Parameter Estimation

Joint likelihood for selection hidden Markov model

The likelihood function for selection hidden Markov model (SHMM) in section 4.3.2 can be described as

$$
\begin{aligned}
L &= \sum_{S} f(O|S,\alpha,\beta)f(S) \\
&= \sum_{S}\{\prod_{i=1}^{N} f(s_{i1}) * \prod_{t=2}^{T} f(s_{it}|s_{i,t-1}) * \prod_{t=1}^{T} f(o_{it}|s_{it},\alpha,\beta)\} \\
&= \prod_{i=1}^{N}\{\sum_{S} f(s_{i1})f(o_{i1}|s_{i1},\alpha,\beta) * \prod_{t=2}^{T} f(s_{it}|s_{i,t-1}) * f(o_{it}|s_{it},\alpha,\beta)\} \\
&= \prod_{i=1}^{N}\{\sum_{S} \pi_{s1} f(o_{i1}|s_{i1},\alpha,\beta) * \prod_{t=2}^{T} \mathbb{Q}s_{it-1,it} * f(o_{it}|s_{it},\alpha,\beta)\} \\
&= \prod_{i=1}^{N}\Big\{\sum_{S} \pi_{s1}\big\{(f(y_{i1},r_{i1}|s_{i1},\alpha,\beta))^{r_{i1}} * (\int f(y_{i1},r_{i1}|s_{i1},\alpha,\beta)dy_{i1})^{1-r_{i1}}\big\} \\
&\times \prod_{t=2}^{T} \mathbb{Q}s_{it-1,it} * \big\{(f(y_{it},r_{it}|s_{it},\alpha,\beta))^{r_{it}} * (\int f(y_{it},r_{it}|s_{it},\alpha,\beta)dy_{it})^{1-r_{it}}\big\}\Big\} \quad .
\end{aligned}
$$

Two stage pseudo-likelihood procedure

For large $S \in (1,2,\cdots,m)$ hidden states, computation is impractical since it involves $\mathcal{O}(Tm^T)$ operations for each subject i and cannot be calculated directly. Baum et al. (1970), Rabiner (1989), and Welch (2003) proposed a type of EM algorithm known as the *backward-forward* or *Baum-Welch* algorithm to solve the estimation in hidden Markov models with discrete time applications, which enjoys the time complexity $\mathcal{O}(Tm^2)$. We propose a two stage pseudo-likelihood method to achieve computational feasibility with a high degree of efficiency. In stage one, we treat the initial probability π and transition matrix $\mathbb{Q}$ as nuisance parameters to simplify the maximum likelihood as a function of the parameters of interest. We first replace all the nuisance parameters in maximum likelihood directly with

their *Baum-Welch* algorithm estimates to form a pseudo maximum likelihood with lower parameter dimensionality; in the second stage, a direct maximization method can be used to maximize the pseudo-likelihood for the parameters of interest, and we continue to iterate until the parameters converge.

We adopt a step by step Baum et al. (1970) procedure to update the nuisance parameters. First we define the forward variables as

$$\mathfrak{a}_{it}(j) = f(o_{i1}, o_{i2}, \cdots, o_{it}, s_{it} = j), \quad i = 1, \cdots, N; t = 1, \cdots, T; j \in S,$$

which denotes the probability of the partial sequence ending up in state j at time t for a given object i. The forward variables $\mathfrak{a}_{it}(j)$ can be calculated recursively by

$$\begin{aligned} \mathfrak{a}_{i1}(j) &= \pi_{s1}(j) * f(o_{i1}|s_{i1} = j) \\ \mathfrak{a}_{i,t+1}(k) &= \sum_{j=1}^{m} \mathfrak{a}_{it}(j) * q_{jk} f(o_{i,t+1}|s_{i,t+1} = k) \quad , \end{aligned}$$

Finding the likelihood by calculating

$$L = \prod_{i=1}^{n} \sum_{j=1}^{m} \mathfrak{a}_{i,T}(j) \quad , \tag{4.1}$$

we define the backward variables

$$\mathfrak{b}_{it}(j) = f(o_{i,t+1}, o_{i,t+2}, \cdots, o_{iT}|s_{it} = j), \quad i = 1, \cdots, N; t = 1, \cdots, T; j \in S,$$

which denotes the probability of the partial sequence in state j at time t from $t+1$ to the end for a given subject i. The backward variables $\mathfrak{b}_{it}(j)$ can be calculated recursively by

$$\begin{aligned} \mathfrak{b}_{iT}(j) &= 1 \\ \mathfrak{b}_{i,t}(j) &= \sum_{k=1}^{m} q_{jk} * f(o_{i,t+1}|s_{i,t+1} = k) * \mathfrak{b}_{i,t+1}(k) \quad . \end{aligned}$$

Define $\hat{\mu}_{itj}$, and $\hat{\nu}_{itj}$ as

$$\begin{aligned}\hat{\mu}_{itj} &= P(S_{it} = j|o_{i,1}, o_{i,2}, \cdots, o_{iT}) \\ \hat{\nu}_{itjk} &= P(S_{it} = j, S_{i,t+1} = k|o_{i,1}, o_{i,2}, \cdots, o_{i,T}) \quad .\end{aligned}$$

Then $\hat{\mu}_{itj}$ and $\hat{\nu}_{itj}$ can be updated using

$$\begin{aligned}\hat{\mu}_{itj} &= \frac{\mathfrak{a}_{it}(j)\mathfrak{b}_{it}(j)}{\sum_{j=1}^{m}\mathfrak{a}_{it}(j)\mathfrak{b}_{it}(j)} \\ \hat{\nu}_{itjk} &= \frac{\mathfrak{a}_{i,t}(j)q_{jk}f(o_{i,t+1}|s_{i,t+1}=k)\mathfrak{b}_{i,t+1}(k)}{\sum_{j,k=1}^{m}\mathfrak{a}_{it}(j)q_{jk}f(o_{i,t+1}|s_{i,t+1}=k)\mathfrak{b}_{i,t+1}(k)} \quad .\end{aligned}$$

We update the transition matrix and initial probability with respect to the initial parameters $\alpha^l, \beta^l, \pi_j^l, q_{jk}^l$:

$$\begin{aligned}\hat{\pi_j}^{l+1} &= \frac{\sum_{i=1}^{n}\hat{\mu}_{i1j}^{l}}{n} \\ \hat{q}_{jk}^{l+1} &= \frac{\sum_{i=1}^{n}\sum_{1}^{T}\hat{\nu}_{itjk}^{l}}{\sum_{i=1}^{n}\sum_{1}^{T}\sum_{k=1}^{m}\hat{\nu}_{itjk}^{l}} \quad .\end{aligned}$$

$\hat{\pi_j}^{l+1}$ is the expected frequency in state j at time $t = 1$, and $\hat{q_{jk}}^{l+1}$ is the expected number of transitions from state j to state k divided by the expected number of transitions from state j. Substituting $\hat{\pi_j}^{l+1}$ and $\hat{q_{jk}}^{l+1}$ into the likelihood function (4.1), we have the pseudo-likelihood function

$$\mathbb{PL}(\alpha, \beta) = \prod_{i=1}^{n}\sum_{j=1}^{m}\mathfrak{a}_{i,T}(j|\alpha^l, \beta^l, \hat{\pi_j}^{l+1}, \hat{q}_{jk}^{l+1}) \quad . \tag{4.2}$$

The quasi-Newton method can then be used to maximize the approximate pseudo-likelihood for α^{l+1}, and β^{l+1}, and we continue the iterations until the parameters α and β converge.

Joint likelihood for the shared parameter hidden Markov model

The likelihood function for the shared parameter hidden Markov model(SPHMM) in section 4.3.3 can be described as

$$
\begin{aligned}
L &= \int_b \sum_S f(O|S,b,\alpha,\beta) f(S) f(b|\psi) db \\
&= \int_b \sum_S \{\prod_{i=1}^{N} f(s_{i1}) * \prod_{t=2}^{T} f(s_{it}|s_{i,t-1}) * \prod_{t=1}^{T} f(o_{it}|s_{it},b,\alpha,\beta)\} f(b|\psi) db \\
&= \int_b \prod_{i=1}^{N} \{\sum_S f(s_{i1}) f(o_{i1}|s_{i1},b,\alpha,\beta) * \prod_{t=2}^{T} f(s_{it}|s_{i,t-1}) * f(o_{it}|s_{it},b,\alpha,\beta)\} f(b|\psi) db \\
&= \int_b \prod_{i=1}^{N} \{\sum_S \pi_{s1} f(o_{i1}|s_{i1},b,\alpha,\beta) * \prod_{t=2}^{T} \mathbb{Q} s_{it-1,it} * f(o_{it}|s_{it},b,\alpha,\beta)\} f(b|\psi) db \\
&= \int_b \prod_{i=1}^{N} \Big\{ \sum_S \pi_{s1} \big\{ (f(y_{i1}|s_{i1},b,\alpha) * f(r_{i1}|s_{i1},b,\beta))^{r_{i1}} * (f(r_{i1}|s_{i1},b,\beta))^{1-r_{i1}} \big\} \\
&\times \prod_{t=2}^{T} \mathbb{Q} s_{it-1,it} * \big\{ (f(y_{it}|s_{it},b,\alpha) * f(r_{it}|s_{it},b,\beta))^{r_{it}} * (f(r_{it}|s_{it},b,\beta))^{1-r_{it}} \big\} \Big\} \\
&\times f(b|\psi) db \quad .
\end{aligned}
$$

For a simple random effects model, considering only one random effect b_i associated with the ith subject ($i = 1, \cdots, N$), assume b_i follows i.i.d. normal distribution. Then, assessments are independent given the sequences of hidden states s_{it} and the random effect b_i. The

one-dimensional random effect likelihood function can be simplified further as

$$\begin{aligned}
L &= \int_b \sum_S f(O|S,b,\alpha,\beta) f(S) f(b|\psi) db \\
&= \int_{b_i} \sum_S \{\prod_{i=1}^{N} f(s_{i1}) * \prod_{t=2}^{T} f(s_{it}|s_{i,t-1}) * \prod_{t=1}^{T} f(o_{it}|s_{it},b_i,\alpha,\beta)\} f(b_i|\psi) db_i \\
&= \prod_{i=1}^{N} \int_{b_i} \Big\{ \sum_S \pi_{s1} \big\{ (f(y_{i1}|s_{i1},b_i,\alpha) * f(r_{i1}|s_{i1},b_i,\beta))^{r_{i1}} * (f(r_{i1}|s_{i1},b_i,\beta))^{1-r_{i1}} \big\} \\
&\times \prod_{t=2}^{T} \mathbb{Q} s_{it-1,it} * \big\{ (f(y_{it}|s_{it},b_i,\alpha) * f(r_{it}|s_{it},b_i,\beta))^{r_{it}} * (f(r_{it}|s_{it},b_i,\beta))^{1-r_{it}} \big\} \Big\} \\
&\times f(b_i|\psi) db_i \quad .
\end{aligned}$$

As in the previous section, forward and backward variables could help in evaluating the likelihood function above and in obtaining parameter estimates. However, for a multi-dimensional random effects model, the forward-backward algorithm is not appropriate since it involves multi-dimensional integration.

A two stage pseudo likelihood procedure as described as section 4.3.4 is used to achieve computational convenience with a high degree of efficiency. Again, first we define the forward variables as

$$\mathfrak{a}_{it}(j,b_i) = f(o_{i1}, o_{i2}, \cdots, o_{it}, S_{it} = j, |b_i), \quad i = 1, \cdots, N; t = 1, \cdots, T; j \in S,$$

which denote the probability of the partial sequence ending up in state j at time t for a given subject i. The forward variables $\mathfrak{a}_{it}(j,b_i)$ can be calculated recursively by

$$\begin{aligned}
\mathfrak{a}_{i1}(j,b_i) &= \pi_j * f(o_{i1}|s_{i1} = j, b_i) \\
\mathfrak{a}_{i,t+1}(k,b_i) &= \sum_{j=1}^{m} \mathfrak{a}_{it}(j,b_i) * q_{jk} f(o_{i,t+1}|s_{i,t+1} = k, b_i) \quad ,
\end{aligned}$$

leading to the likelihood

$$L = \prod_{i=1}^{n} \int_{b_i} \sum_{j=1}^{m} \alpha_{i,T}(j, b_i) h(b_i|\psi) db_i, \tag{4.3}$$

where $h(\cdot|\psi)$ is the density function of b_i. Second, we define the backward variables

$$\mathfrak{b}_{it}(j, b_i) = f(o_{i,t+1}, o_{i,t+2}, \cdots, o_{iT}|S_{it} = j, b_i), \quad i = 1, \cdots, N; t = 1, \cdots, T; j \in S \quad ,$$

which denote the probability of the partial sequence starting in state j at time t from $t+1$ to the end for a given object i. The backward variables $\mathfrak{b}_{it}(j, b_i)$ can be calculated recursively by

$$\begin{aligned} \mathfrak{b}_{iT}(j, b_i) &= 1 \\ \mathfrak{b}_{i,t}(j, b_i) &= \sum_{k=1}^{m} q_{jk} * f(o_{i,t+1}|s_{i,t+1} = k, b_i) * \mathfrak{b}_{i,t+1}(k, b_i) \quad . \end{aligned}$$

Define $\hat{\mu}_{itj}$, and $\hat{\nu}_{itj}$ as

$$\begin{aligned} \hat{\mu}_{itj} &= P(S_{it} = j|o_{i,1}, o_{i,2}, \cdots, o_{iT}) \\ \hat{\nu}_{itjk} &= P(S_{it} = j, S_{i,t+1} = k|o_{i,1}, o_{i,2}, \cdots, o_{i,T}) \quad . \end{aligned}$$

These can be calculated directly by

$$\begin{aligned} \hat{\mu}_{itj} &= \frac{\int \mathfrak{a}_{it}(j, b_i) \mathfrak{b}_{it}(j, b_i) h(b_i) db_i}{\int \sum_{j=1}^{m} \mathfrak{a}_{it}(j, b_i) \mathfrak{b}_{it}(j, b_i) h(b_i|\psi) db_i} \\ \hat{\nu}_{itjk} &= \frac{\int \mathfrak{a}_{it}(j, b_i) q_{jk} f(o_{i,t+1}|S_{i,t+1} = k, b_i) \mathfrak{b}_{i,t+1}(k, b_i) h(b_i|\psi) db_i}{\int \sum_{j,k=1}^{m} \mathfrak{a}_{it}(j, b_i) q_{jk} f(o_{i,t+1}|S_{i,t+1} = k, b_i) \mathfrak{b}_{i,t+1}(k, b_i) h(b_i|\psi) db_i} \quad . \end{aligned}$$

We then update the transition matrix and initial probability with respect to the initial

parameters $\alpha^l, \beta^l, \pi_j^l, q_{jk}^l, \psi^l$:

$$\begin{aligned} \hat{\pi_j}^{l+1} &= \frac{\sum_{i=1}^n \hat{\mu}_{i1j}^l}{n} \\ \hat{q_{jk}}^{l+1} &= \frac{\sum_{i=1}^n \sum_1^T \hat{\nu}_{itjk}^l}{\sum_{i=1}^n \sum_1^T \sum_{k=1}^m \hat{\nu}_{itjk}^l} \quad . \end{aligned}$$

Substituting $\hat{\pi_j}^{l+1}$ and $\hat{q_{jk}}^{l+1}$ into likelihood function (4.3), we have the pseudo-likelihood function

$$\mathbb{PL}(\alpha, \beta, \psi) = \prod_{i=1}^n \int_{b_i} \sum_{j=1}^m \mathfrak{a}_{i,T}(j, b_i | \alpha^l, \beta^l, \psi^l, \hat{\pi}_j^{l+1}, \hat{q}_{jk}^{l+1}) h(b_i|\psi) db_i \quad . \tag{4.4}$$

Quasi-Newton methods can then be used to maximize the approximate pseudo-likelihood for α^{l+1}, β^{l+1}, and ψ^{l+1}. We continue the iteration until the parameters α, β, and ψ converge.

Variance-covariance estimation

The pseudo-score function is defined as

$$\mathcal{S}_T(\alpha, \beta) = \sum_{i=1}^n \mathcal{S}_{Ti}(\alpha, \beta) = \frac{\partial}{\partial(\alpha, \beta)} \log \mathbb{PL}_i \quad ,$$

and the maximum pseudo-likelihood estimate is the solution to $\mathcal{S}_T(\hat{\alpha}, \hat{\beta}) = \mathbf{0}$. Heuristically, using method of moments ideas, the pseudo-score estimator is consistent if the distributions $f(y_{it}, r_{it}|\mathbf{X_i}, \mathbf{S_{it}}, \alpha, \beta)$ (SHMM), and $f(y_{it}, r_{it}|\mathbf{X_i}, \mathbf{S_{it}}, \alpha, \beta, \mathbf{b_i})$ (SPHMM) are correctly specified. Troxel et al. (1998b) gave proof of the consistency of the pseudo-likelihood estimator. It can be shown that $\mathbf{E}[\mathcal{S}_T(\alpha, \beta)] = 0$ at the true (α, β). In practice, we obtain $(\hat{\alpha}, \hat{\beta})$ by maximizing the log-pseudolikelihood directly, but the solution satisfies $\mathcal{S}_T(\hat{\alpha}, \hat{\beta}) = \mathbf{0}$. The variances have to be adjusted to obtain correct inference because of the assumptions about the transition matrix. We accomplish this with the commonly-used sandwich estimator as

in Liang and Zeger (1986):

$$\Sigma = \Big[\frac{1}{n}E\Big\{\frac{\partial \mathcal{S}_T(\alpha,\beta)}{\partial(\alpha,\beta)}\Big\}\Big]^{-1}\frac{1}{n}\sum_{i=1}^{n}E\{\mathcal{S}_{Ti}(\alpha,\beta)\mathcal{S}'_{Ti}(\alpha,\beta)\}\Big[\frac{1}{n}E\Big\{\frac{\partial \mathcal{S}_T(\alpha,\beta)}{\partial(\alpha,\beta)}\Big\}\Big]^{-1}.$$

Furthermore, the robust variance estimate $\hat{\Sigma}$ is obtained by simply replacing (α,β) by $(\hat{\alpha},\hat{\beta})$ in the above expression.

4.3.5. Numerical integration

There is no closed form for the pseudo-likelihood function (4.2,4.4) due to the joint likelihood; a numerical integration method will be applied to approximate the pseudo-likelihood function (4.2,4.4). Laplacian, Gaussian Quadrature or Adaptive Gaussian Quadrature can be used to approximate the integration numerically for low dimensional shared parameter hidden Markov model (SPHMM) or the one dimensional selection hidden Markov model (SHMM). For Gaussian data, Gaussian quadrature methods offer both accuracy and efficiency. The quasi-Newton Method can then be used to maximize the approximate likelihood. Unlike the EM algorithm, direct maximization of the log-pseudo-likelihood requires good initial values of the parameters. One approach is to choose a vector of starting values and fit a HMM model assuming MCAR to get the starting points as close to the true values as possible. On the other hand, for large numbers of random effects, numerical integration methods are no longer appropriate for SPHMM. Then Monte Carlo expectation-maximization (MCEM) algorithm or simulated maximum likelihood methods (McCulloch, 1997; Jank and Booth, 2003) are more feasible.

4.4. Simulation Study

In this section we define the following simulation study to investigate the empirical behavior of the proposed models. To model continuous observations with Gaussian distribution we generated 500 repeated samples of size $n = 150, 300$ and $T = 3$ according to the following scheme.

4.4.1. SHMM:

$$(y_{it}|S_{it} = j) \sim \mathbf{Normal}(\mu_{itj}, \sigma^2), \quad j = 1, 2$$

where the following mean function holds:

$$\mu_{itj} = \alpha_{j0} + \alpha_{j1}x_{it1} + \alpha_{j2}(t-1)$$

and for the outcome misssingess mechanism model:

$$\mathbf{logit}(\mathbf{Pr}(\mathrm{R_{it}} = 1|\mathrm{Y_{it}})) = \beta_0 + \beta_1 x_{it1} + \beta_2(t-1) + \beta_3 * Y_{it} \quad .$$

The covariates x_{it1} were independently drawn from a Bernoulli distribution with $p = 0.5$ and Y_{it} is the continuous outcome observed at time t on patient i with common standard deviation $\sigma = 0.35$. R_{it} and S_{it} are the associated missingness indicator (1=observed, 0=missing) and hidden state (1=remission, 2=relapse). We assume the following true values for the parameter vectors.

For the nuisance parameter:

$$\pi = \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} = \begin{bmatrix} 0.65 \\ 0.35 \end{bmatrix}, \mathbb{Q} = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} = \begin{bmatrix} 0.40 & 0.60 \\ 0.35 & 0.65 \end{bmatrix}$$

and marginal effects:

$$\alpha = \begin{bmatrix} \alpha_{10} & \alpha_{20} \\ \alpha_{11} & \alpha_{21} \\ \alpha_{12} & \alpha_{22} \end{bmatrix} = \begin{bmatrix} 0.65 & -1.5 \\ 1.05 & 1.55 \\ 0.25 & 0.75 \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 3.45 \\ -0.55 \\ -0.35 \\ -1.55 \end{bmatrix} \quad .$$

4.4.2. SPHMM:

$$\begin{aligned} (y_{it}|S_{it} = j, b_i) &\sim \textbf{Normal}(\mu_{itj}, \sigma^2), \quad j = 1, 2 \\ b_i &\sim \textbf{Normal}(0, \psi^2) \end{aligned}$$

where the following mean function holds:

$$\mu_{itj} = \alpha_{j0} + \alpha_{j1} x_{it1} + \alpha_{j2}(t-1) + b_i$$

and for the misssingess mechanism model:

$$logit(Pr(R_{it} = 1|Y_{it}, b_i)) = \beta_0 + \beta_1 x_{it1} + \beta_2(t-1) + b_i \quad .$$

The random effects b_i are independently drawn from $N(0, 0.85^2)$ and all other parameters are the same as in the SHMM model described in section 4.4.1. The simulation was conducted to assess the behavior of the proposed model with respect to both sample size n and to potential miss-specification, compared to the SHMM defined in (4.3.2) and the SPHMM defined in (4.3.3).

Tables 4.2 and 4.3 give the sample mean, sample standard deviation, average asymptotic 95% confidence interval and coverage probability of the parameter estimates obtained based on 500 simulations from each model. In the left column, we list the comparison of SHMM and SPHMM with models correctly specified. In right column, we list the comparison of SPHMM and SHMM with models miss-specified. The true parameters were selected to

generate a substantial amount of missing response.

When the models are correctly specified, the mean values are close to the true parameter values with high coverage probabilities; the "convenience" estimator of the mean of the nuisance parameters obtained from the *backward-forward* algorithm is close to the true Markov chain parameters as well. Increasing the sample size from $n = 150$ (Tables 4.2) to $n = 300$ (Table 4.3) shows clear improvement in both the marginal model and Markov chain parameters. Both models achieved better mean values, and narrower 95% confidence intervals with higher coverage probability.

When the models are mis-specified, the mean values of the missingness mechanism model for both SHMM and SPHMM are severely biased. The estimated parameters of the missingness mechanism model tend to be overestimated strongly. The parameters in the state dependent model become less efficient. However, SPHMM still provides much better estimators of both state-dependent model and Markov chain parameters than SHMM. This is not surprising since the random effect introduced in SPHMM provides more flexibility, and efficiency, and relaxes the assumption that the observations are conditionally independent given the hidden states, especially assuming hidden states as a category number. Simulations show that the random effect b_i introduced in SPHMM model handles the mis-specified situation better, since the random effect itself absorbs the potential extra effects. However, SPHMM takes substantially longer computational time than SHMM in this small simulation study.

4.4.3. Sensitivity analysis to transition matrix

The estimation of the marginal model suffers the issue of power loss when nuisance parameters are mis-specified (Gong and Samaniego (1981),Guolo (2011)). The inefficiency tends to be more severe when the pre-specified transition matrix departs from the true underlying structure. Simulations in this section follow the two schemes below with $n = 500$ sample size and 500 replications. Conclusions are similar for both SHMM and SPHMM; we present results only for the SHMM.

Scheme $\mathbb{Q}_A$ has the true data generated considering a model for disease progression with no recovery, and with the transition matrix as a fully connected or ergodic structure. Notice that the SHMM model will estimate zero using q_{21}.

$$\mathbb{Q}_A = \begin{bmatrix} 0.40 & 0.60 \\ 0 & 1 \end{bmatrix}_{\textbf{True}}, \quad \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix}_{\textbf{Fitted}}$$

Scheme $\mathbb{Q}_B$ has the true data generated considering fully connected or ergodic SHMM model, and fitted the transition matrix as a pre-specified disease progression model with no recovery. Notice that the SHMM will fix the nuisance parameters $q_{12} = 0$ and $q_{22} = 1$. It is clear this has no effect on the re-estimation procedure since any SHMM parameters set to zero initially will remain at zero throughout.

$$\mathbb{Q}_B = \begin{bmatrix} 0.40 & 0.60 \\ 0.35 & 0.65 \end{bmatrix}_{\textbf{True}}, \quad \begin{bmatrix} q_{11} & q_{12} \\ 0 & 1 \end{bmatrix}_{\textbf{Fitted}}$$

Tables 4.4 and 4.5 give the sample mean, sample standard deviation, average asymptotic 95% confidence interval and coverage probability of the parameter estimates obtained based on 500 simulations from each transition matrix scheme. Clearly, these estimators become inefficient due to the mis-specified transition matrix. However, the fully connected transition structure Q_A in Table 4.4 provides more robust and flexible estimation than the strictly constrained structure Q_B in Table 4.5. One should exercise caution when introducing zeroes into the transition matrix; although it reduces the parameter space, it increases inefficiency and leads to severely biased estimators.

4.5. Example: Analysis of RTOG Data

There were 289 eligible participants aged 18 years or older with newly diagnosed anaplastic oligodendroglioma (AO) brain tumors. The eligibility criteria for RTOG 9402 were previously described by Cairncross et al. (2006) and Wang et al. (2010). Eligible participants

were randomized to either procarbazine, lomustine, and vincristine (PCV) chemotherapy plus radiation (arm 1) or radiation alone (arm 2). Patients had to begin the treatment within 1 week of randomization. The chemotherapy regimen used in this study was intensive PCV (I-PCV) which is 25% stronger than standard PCV. I-PCV was given in four week cycles every six weeks followed by radiation. The radiation regimen used in this study was external beam RT 59.4 Gy (1.8 Gy x 33 fractions, 5 days a week) to MR defined tumor volume; radiation was given soon after surgery in arm 2 (within 8 weeks of diagnosis). Patients were stratified by age (younger than 50 years vs over 50 years), Karnofsky performance status (KPS) of 60-70 vs 80-100, and anaplastic tumor grade (2-3 vs 4-5). Table 4.1 gives the patients characteristics by each arm.

The mini-mental status exam (MMSE) is a well known tool used to assess mental status. It is an 11-question measurement that tests five areas of cognitive function: orientation, registration, and repetition; complex commands; attention and calculation; recall; and language. MMSE scores range from 0 to 30 points. A score of 25 or lower indicates a cognitive abnormality. The Quality of Life Questionnaire (QLQ) was developed by the European Organization for Research and Treatment of Cancer (EORTC) to assess the impact of cancer and its treatment on patients' lives. The B-QLQ, modified and developed by Mackworth (1992) to apply to brain cancer patients, was used in RTOG 9402 to evaluate patients' global quality of life and emotional well-being. This is a 100 point scale. Higher QLQ scores, suggest better the quality of life. The MMSE form was completed by the nurse, research associate, or physician, reflecting the opinion of the health care specialist; the B-QLQ was reported by patients themselves, reflecting the patients' point of view. The MMSE and B-QLQ were assessed at baseline and each follow-up visit and then at yearly intervals until the end of follow up.

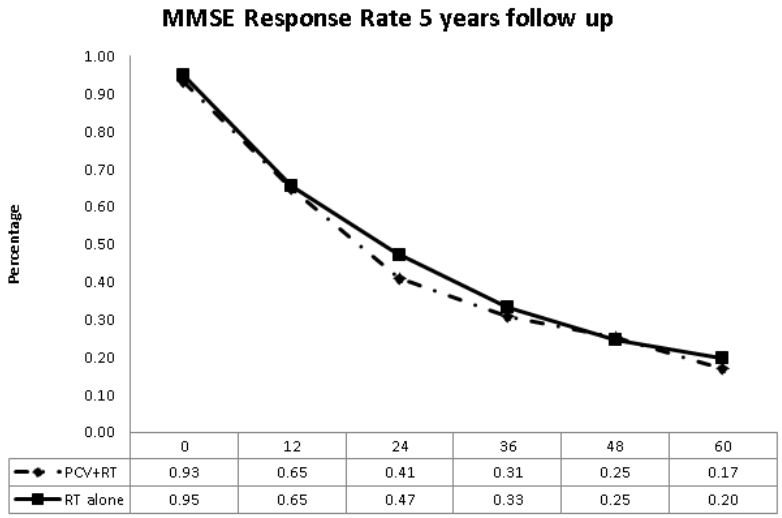
Previous reports on RTOG 9402 showed that patients who have the 1p and 19q chromosomes deletion had longer progression free survival times, but also substantial toxicity in PCV+RT arm. Median survival time was improved in participants in the PCV+RT arm as opposed

to the RT only arm (14.7 years vs. 7.3 years). There was no significant difference in median survival times between the two treatment arms in patients with only one deletion or no deletions of chromosomal segment. In this article, we focus on the association between patients' MMSE/B-QLQ scores and treatment effect. The MMSE and B-QLQ scores are the primary outcomes.

4.5.1. Data analysis: 5 year followup with full data

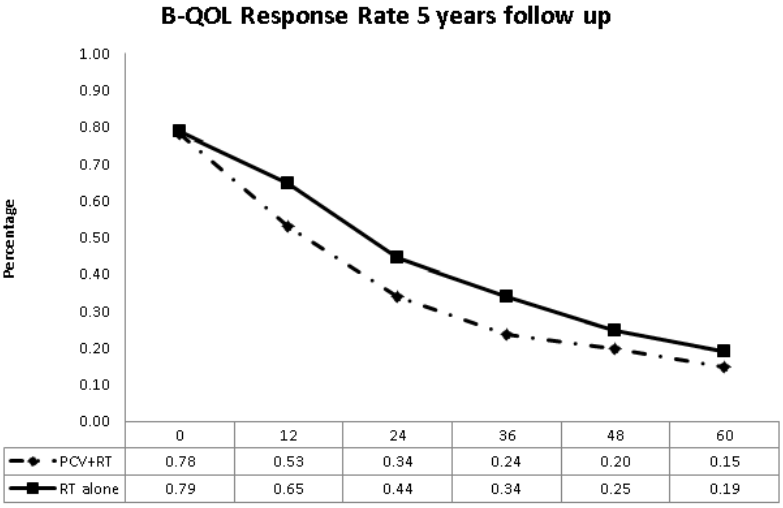
In the first analysis, we include all 289 patients in the cohort. The missingness mechanism model models the overall probability of response to the MMSE/B-QLQ; we do not distinguish between dropout due to death and dropout due to other reasons in order to take advantage of the full sample size by including all patients who entered the trial. In reality, patients who died probably differ systematically from patients who dropped out; to address this, we conducted a second data analysis to evaluate the treatment effect in subjects who survived at least two years, presented in Section 5.2. All models are estimated assuming two hidden states ($S = 2$) due to the relatively limited sample size.

The outcomes MMSE and B-QLQ scores are highly skewed. We use a logarithm transformation for both outcomes to reduce the skewness. Figures 4.1 and 4.2 show the response rates for MMSE and B-QLQ scores over the full five years of follow-up. Non-response includes intermittent missing data, dropout (i.e., study withdrawal), and death. There are total of 101 (35%) patients who died during the 5 year follow-up; 111 (38%) patients dropped out due to unknown reasons. For the MMSE, there are 41 (14%) patients who have at least one intermittent missing value; only 29 (10%) completed all assessments. For the B-QLQ, 44 (15%) patients have at least one intermittent missing value; only 33 (11%) patients finished all questionnaires. Patients who never completed a questionnaire are excluded from the respective analyses.



	0	12	24	36	48	60
PCV+RT	0.93	0.65	0.41	0.31	0.25	0.17
RT alone	0.95	0.65	0.47	0.33	0.25	0.20

Figure 4.1: MMSE response rate for 5 years follow up.



	0	12	24	36	48	60
PCV+RT	0.78	0.53	0.34	0.24	0.20	0.15
RT alone	0.79	0.65	0.44	0.34	0.25	0.19

Figure 4.2: B-QLQ response rate for 5 years follow up.

MMSE outcomes

The parameters are estimated using the proposed methods, comparing SPHMM and SHMM in Table 4.6. The estimators are consistent if the model are correctly specified. SPHMM and SHMM differ in how they relate the probability of the response process and the missingness mechanism. SPHMM links the two by relating a subject's outcome value to the propensity to missingness, whereas SHMM directly models the probability of missingness as a function

of the response. The choice of a modeling framework may depend on the data generating process. Longitudinal data in which missingness is believed to be related to the disease process and not to a particular realization of the outcome may be more appropriately modeled by SPHMM than SHMM. In addition, the simulation study indicated that the SPHMM performs better and is robust when the model is misspecified.

In the missingness models for SPHMM and SHMM, the parameter σ_{b_i} (0.071, $p < 0.001$) and the coefficient of Y_i (5.805, $p < 0.001$), respectively, are significant, indicating that the missingness is "non-ignorable." In the SPHMM, the initial probability $\pi_1 = 0.972$ for state 1 suggests that the hidden states S related to MMSE are most likely very homogeneous, that is, the initial rating of cognitive ability by health care specialists are all very similar at beginning of the trial. State 1 is likely a "stable" state: patients in state 1 have better MMSE scores, and the PCV+RT arm does not significantly affect patients' cognitive ability compared to RT alone (0.000, $p = 0.908$). Only KPS level and age affect cognitive ability. Patients with better KPS levels have better MMSE scores (0.047, $p < 0.001$), and older (50 years plus) patients have worse MMSE scores than younger patients (50 years under) (-0.038, $p = 0.002$). However, there are a few patients falling in state 2 ($\pi_2 = 0.028$), which is more likely a "responding" state. The patients in state 2 have lower initial MMSE scores than patients in state 1. They do respond to the PCV+RT treatment (0.698, $p < 0.001$). Patients in stage 2 who had total resection surgery do worse than patients who only had biopsy or partial resection before treatment (-0.568, $p < 0.001$); KPS level and age do not affect patients' MMSE scores while in state 2 (the coefficients are not statistically significant).

In the missingness mechanism model, the assessment time (-0.768, $p < 0.001$), KPS level (0.452, $p = 0.026$) and patients' age (-0.586, $p = 0.003$) are statistically significant. Patients tend to respond less as time increases, patients who have better KPS level tend to have better MMSE scores, and younger patients (50 years under) tend to be more responsive to the MMSE survey then older patients. The SHMM model is consistent with the

SPHMM model for covariates with significant effects. However the result from SHMM may be overestimated. The p-value were calculated from Wald statistics.

B-QLQ outcomes

In Table 4.7, we list the parameter estimates for the B-QLQ scores. B-QLQ scores are patients' self-report scores reflecting the impact of disease and treatment during the study. The treatment may improve patients' survival time but reduce quality of life dramatically, especially for patients in Arm 1 with intensive PCV chemotherapy.

The parameter σ_{b_i} (0.194, $p < 0.001$) and the coefficient of Y_i (5.671, $p < 0.001$) in the SPHMM and SHMM, respectively, are statistically significant which indicates that the missingness is "non-ignorable." In the SPHMM, the initial probability $\pi_1 = 0.789$ for state 1 shows the initial reporting of quality of life can be separated into two states. The state 1 is more likely a "deteriorating" stage. Patients in state 1 have slightly lower initial B-QLQ scores than patients in state 2. There were no significantly different treatment effects on the study groups for patients in either state, $(-0.044, p = 0.355)$ for state 1 and $(0.052, p = 0.310)$ for state 2. The assessment time, and KPS level affect the B-QLQ scores for patients in state 1. Patients with better KPS levels (0.130, $p = 0.008$) experienced better B-QLQ, and worse B-QLQ as time increase (-0.029, $p = 0.017$). State 2 is more likely a "stable" state with slightly better initial B-QLQ scores. KPS level and age affect the B-QLQ scores for patients in state 2. Patients with better KPS levels have better B-QLQ scores (0.164, $p = 0.018$); older (50+) patients experienced worse B-QLQ scores than younger patients (50−) -0.153, $p = 0.004$. This is similar to what we saw for the MMSE outcome. The transition probabilities $q_{12} = 0.499$ and $q_{21} = 0.489$ indicate that these self-reported B-QLQ scores were quite variable and move often between states. The B-QLQ scores may be subject to patients' mood or other unmeasured characteristics at the time.

In the missingness mechanism model, the assessment time (-0.702, $p < 0.001$), KPS level (0.532, $p = 0.009$) and patients' age (-0.587, $p = 0.003$) are statistically significant. Pa-

tients tend to respond less as time increase, patients who have better KPS levels tend to respond to the B-QLQ better, and younger patients (50 years under) tend to be more responsive to the B-QLQ than older patients (50 years plus). The coefficient of PCV+RT (-0.322, $p = 0.092$) did not achieve statistical significance although the estimate suggestion a negative effect. The SHMM model is consistent with the SPHMM model for covariates with significant effect.

4.5.2. Data analysis: subject with at least 2 years of follow-up.

In the second analysis, we restricted the cohort to patients who survived to at least 2 years; most patients who were excluded died within first year in this study. The results show that patients who died experienced much worse cognitive ability and worse quality of life. The outcomes (MMSE and B-QLQ) may not truly reflect the treatment effect in patients with such short-term survival. Table 4.8 gives the patients characteristics by arm for the restricted cohort. There are 201 (69.55%) patients included in this study cohort. The proportion of subjects in each arm is similar to that in Table 4.1

Figures 4.3 and 4.4 give the plots for response rates of the MMSE and the B-QLQ scores of patients who survived at least 2 years.

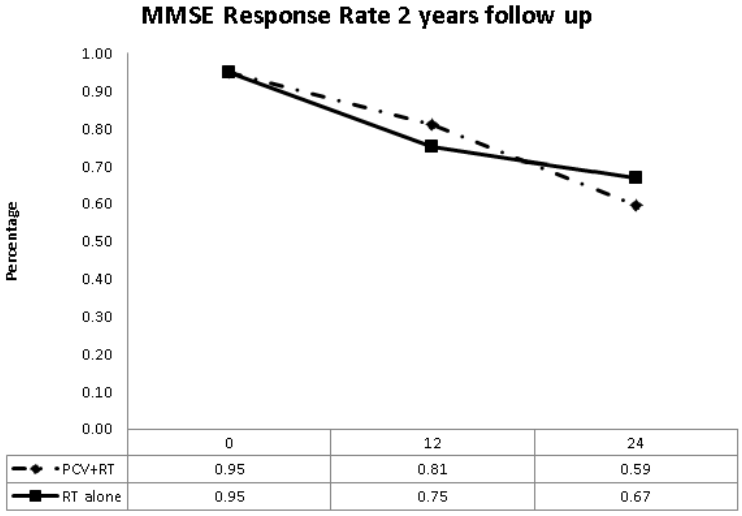


Figure 4.3: MMSE: response rate for at least 2 years survival.

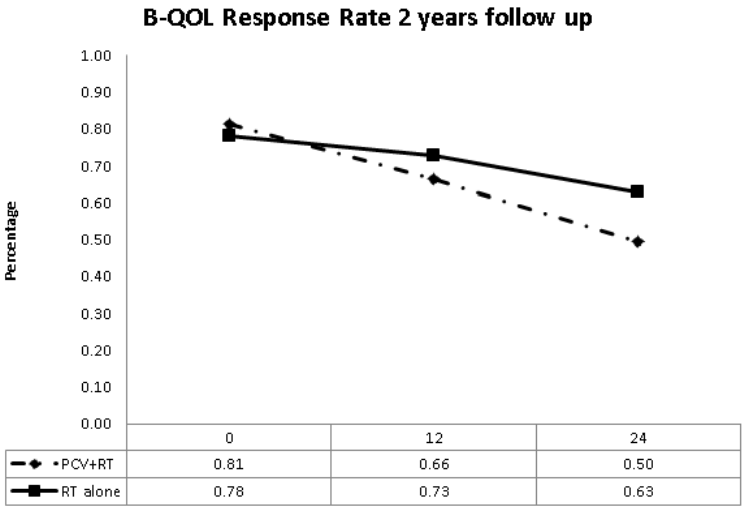


Figure 4.4: B-QOL: response rate for at least 2 years survival.

MMSE outcomes

In Table 4.9, we list the parameter estimates for the MMSE scores after limiting the cohort to those surviving at least 2 years. The parameter σ_{b_i} (0.029, $p < 0.001$) and the oefficient of Y_i (4.459, $p < 0.001$) in the SPHMM and SHMM, respectively, are statistically significant which indicates the missingness is "non-ignorable." In the SPHMM, the initial probability $\pi_1 = 0.984$ for state 1 is similar to the full data analysis. The initial rating of cognitive ability by health care specialists most likely reflect one state. State 1 is a "stable" state. Patients in state 1 have better MMSE scores, and the PCV+RT arm does not significantly affect patients cognitive ability compared to RT alone (0.003, $p = 0.708$). The resection, KPS level and age have statistically significant effects on patients' cognitive ability. Patients with better KPS levels do better (0.044, $p < 0.001$); those older than 50 years have worse MMSE scores than younger patients (50−) (−0.024, $p = 0.011$); and patents undergoing total resection experience worse MMSE scores (−0.016, $p = 0.045$). Similarly, there are few patients starting in state 2 $\pi_2 = 0.016$, which is more likely a "responding" state. Patients in state 2 have lower initial MMSE scores than patients in state 1. The treatment effect we saw in the full data analysis is no longer statistically significant (0.102, $p = 0.079$).

However, the coefficient of the assessment time is positively associated with patients' cognitive ability (0.234, $p < 0.001$). This is consistent with the data. We see increased MMSE scores on average in first the two years among surviving patients. The patients undergoing total resection have worse MMSE scores than those undergoing biopsy or partial resection ($-0.127, p < 0.013$). The resection effects are statistically significant in both states. Patients' age (-0.2, $p = 0.011$) and grade (0.309, $p < 0.001$) are statistically significant. Older (50+) patients have worse MMSE scores than younger patients (50−) (-0.02, $p = 0.011$). The effect of grade reflects the trend which we saw in the data as well. There were increased MMSE scores on average comparing severe grade $4-5$ to moderate grade $2-3$. This suggests that the patients who initially had poor levels of cognitive ability actually show more improvement in cognitive ability in the first two years. In the missingness mechanism model, only the assessment time (-1.317, $p < 0.001$) is statistically significant. Patients tend to respond less as time increase. The SHMM model is generally consistent with the SPHMM model for effects in the outcome model.

B-QLQ outcomes

In Table 4.10, we list the parameter estimates for the B-QLQ scores after limiting the cohort to those who survive at least 2 years. The parameter σ_{b_i} (0.188, $p < 0.001$) and the coefficient of Y_i (4.459, $p < 0.001$) in the SPHMM and SHMM, respectively, are statistically significant which indicates that the missingness is "non-ignorable." In the SPHMM, the initial probability $\pi_1 = 0.875$ for state 1, which is again likely a "stable" stage. Patients in state 1 have higher initial B-QLQ scores, and PCV+RT arm does not significantly affect patients' B-QLQ scores compared to RT alone (-0.043, $p = 0.215$). Only KPS level affect the B-QLQ scores. Patients with better KPS levels experienced better B-QOL scores (0.116, $p = 0.002$). State 2 is more likely a "deteriorating" stage with a worse initial B-QLQ scores than those in state 1. The assessment time, PCV+RT, resection, KPS level and grade are all significantly associated with the B-QLQ score in state 2. Patients with better KPS levels have better B-QLQ scores (0.734, $p < 0.001$); patients experienced worse B-QLQ as

time passes (-0.177, $p = 0.008$); patients in the PCV+RT arm had worse B-QLQ scores (-0.240, $p = 0.021$) than patients receiving RT alone; patients undergoing total resection had better B-QLQ scores (0.229, $p = 0.031$); patients with severe grade ($4-5$) experienced worse B-QLQ scores than patients with moderate grade ($2-3$) (-0.268, $p = 0.001$).

In the missingness mechanism model, assessment time (-0.811, $p < 0.001$) and grade (0.550, $p = 0.033$) are statistically significant. Patients tend to be less responsive with longer follow up, and patients who have severe grade initially tend to respond to the B-QLQ form more often. Treatment arm does not significantly affect patients' decision to respond to the B-QLQ questionnaire (-0.379, $p = 0.122$).

Summary

We see some differences after excluding patients with short-term survival. This is what we expected, since these patients have generally worse MMSE and B-QLQ scores. For the MMSE scores, the treatment effect become less significant in state 2 after we excluded those short-term survival patients; this makes intuitive sense since the restricted cohort does not include as many patients with very poor cognitive function who have room for considerable improvement. Those patients undergoing total resection had worse cognitive ability. There is a suggestion that patients with initial worse health status respond better to treatment than those patients with better baseline health status. For this restricted cohort, our model results are consistent with the empiric data. We did see increased MMSE scores on average as time passes, and increased MMSE scores on average comparing severe grade ($4-5$) vs moderate grade ($2-3$). For the B-QLQ scores, the treatment effect become more significant in state 2 after we excluded the short-term survival patients; this may reflect the room for improvement in the small subset who start out with lower quality of life. Patients in state 2 undergoing the PCV+RT treatment have statistically significantly lower B-QLQ scores. From the health specialists' point of view, radiation treatment is not expected to improve cognitive ability. On the other hand, patients in the PCV+RT arm

experienced decreasing quality of life, if they had poor initial health status. This may reflect the different expectations of heath specialist and patients themselves about the potential effects of chemotherapy. All the models we have seen suggest "non-ignorable" missing data, that is, subjects with poor outcomes are more likely to have missing values. Thus it is critical to treat the missingness model correctly in order to achieve valid estimates of the effects of interest.

4.6. Discussion

We have presented an extension of a pseudo likelihood-based algorithm to handle "non-monotone" and "non-ignorable" missing data. We assumed a hidden Markov structure, which is a natural way to capture the changes in outcomes among repeated measurements in a longitudinal data setting. The conditional independence assumed in the hidden Markov model provides a simple framework for reducing the multi-dimensional integration in traditional methods into one dimensional integration in the observed likelihood. In addition, the proposed models avoid the problem of specification of the correlation structure of repeated outcomes. By modeling the outcome progression through different hidden states, our approach gives more targeted estimates of the covariate effects.

Our transition model can be easily extended to models with more than two states, such as dropout or intermittent missingness. The numerical integration provides an accurate approximation but at the cost of increased computational complexity. Direct maximization of the log-pseudo-likelihood, as used here, requires good initial values of the parameters. One approach is to choose a vector of starting values and use GEE estimates to get the starting points as close as the true values as possible. The main effects and missingness effects are consistent with high coverage probabilities as long as the models are correctly specified. Increasing the sample size will help to stabilize the estimation of the initial probability of each hidden state, and increasing the number of assessment times will facilitate estimation of the transition matrix. Derived from theory of pseudo likelihood-based methods, the proposed method requires a large sample size to perform better. In the shared parameter model, the

normal assumption on both the outcomes and the random effect seems questionable, especially considering the highly skewed distribution of outcomes in our example. A Weibull model with Gamma random effects (Chen et al., 2009) may be better suited to such highly skewed longitudinal data. The distribution of the random effect assumed here may cause some sensitivity in our result due to lack of information in the data. It is also possible to extend this method to account for time effects in the Markov model. The optimal number of hidden states can be selected based on AIC/BIC criterion. MacKay (2002) gives a discussion and an alternative model selection criterion for the simple hidden Markov model. Shared parameter models and selection models are different in how they relate the outcome process and the missingness mechanism. Shared parameter models link the two by relating a subject's outcome to the propensity for missingness; and selection models directly model the probability of missingness as a function of the outcome. So the choice of a modeling framework may depend on the data generating process. Longitudinal data in which missingness is believed to be related to the disease process and not to a particular realization of this process may be more appropriately modeled by a shared parameter model than a selection model.

As with any model-based approach to non-ignorable missing data, the current approach is subject to unavoidable assumptions about the complete data distribution and the missing data mechanism. It is important to consider all substantive information about the area of application, prior experience with missing data in similar situations, and expert opinion about the mechanism of missing data when building such models. In many areas, enough knowledge and experience exists to justify the necessary assumptions, and the benefit in terms of bias reduction can be significant.

Table 4.1: Patients characteristics by arm

		PCV+RT		RT		
			%		%	p-value
		147	50.87	142	49.13	
Age						0.956
	50 >	101	68.71	98	69.01	
	50 <	46	31.29	44	30.99	
Resection						0.462
	biopsy/partial	62	42.18	66	46.48	
	total resection	85	57.82	76	53.52	
KPS						0.404
	60–80	41	27.89	46	32.39	
	90–100	106	72.11	96	67.61	
Grade						0.743
	anaplastic (2-3 features)	80	54.42	80	56.34	
	anaplastic (4-5 features)	67	45.58	62	43.66	

Table 4.2: Simulation of model comparison $n = 150$: SHMM vs SPHMM

n=150										
Rsp.Rate	$T_1 = 0.817$ $T_2 = 0.776$ $T_3 = 0.617$									
		SHMM Correctly specified					SPHMM Mis-specified			
	True	Est	95% L	95% U	C.P		Est	95% L	95% U	C.P
Parameters						Parameters				
α_{10}	0.650	0.650	0.556	0.745	0.928	α_{10}	0.642	0.550	0.734	0.924
α_{11}	1.050	1.050	0.885	1.215	0.946	α_{11}	0.956	0.816	1.097	0.726
α_{12}	0.250	0.253	0.149	0.358	0.934	α_{12}	0.223	0.136	0.311	0.878
α_{20}	-1.500	-1.500	-1.602	-1.398	0.916	α_{20}	-1.482	-1.583	-1.382	0.918
α_{21}	1.550	1.546	1.431	1.660	0.962	α_{21}	1.492	1.386	1.599	0.820
α_{22}	0.750	0.750	0.677	0.823	0.932	α_{22}	0.729	0.660	0.798	0.902
σ	0.350	0.345	0.292	0.398	0.914	σ	0.335	0.278	0.392	0.836
Missingness mechanism										
β_0	3.450	3.556	2.317	4.796	0.962	β_0	3.030	2.412	3.649	0.684
β_1	-0.550	-0.549	-1.801	0.704	0.968	β_1	-2.112	-2.653	-1.571	0.000
β_2	-0.350	-0.342	-0.700	0.017	0.964	β_2	-0.634	-0.936	-0.332	0.540
β_3	-1.550	-1.597	-2.887	-0.308	0.972	ψ	0.009	-0.323	0.341	0.480
Nuisance parameter										
Markov Chain parameters										
π						π				
π_1	0.650	0.654				π_1	0.592			
π_2	0.350	0.346				π_2	0.408			
$\mathbb{Q}$						$\mathbb{Q}$				
q_{11}	0.400	0.383				q_{11}	0.310			
q_{12}	0.600	0.617				q_{12}	0.690			
q_{21}	0.350	0.374				q_{21}	0.294			
q_{22}	0.650	0.626				q_{22}	0.706			

n=150										
Rsp.Rate	$T_1 = 0.945$ $T_2 = 0.803$ $T_3 = 0.516$									
		SPHMM Correctly specified					SHMM Mis-specified			
	True	Est	95% L	95% U	C.P		Est	95% L	95% U	C.P
Parameters						Parameters				
α_{10}	0.650	0.664	0.432	0.897	0.944	α_{10}	0.607	0.279	0.936	0.892
α_{11}	1.050	1.045	0.723	1.367	0.944	α_{11}	1.156	0.714	1.598	0.892
α_{12}	0.250	0.242	0.128	0.357	0.928	α_{12}	0.191	0.036	0.345	0.842
α_{20}	-1.500	-1.496	-1.751	-1.242	0.930	α_{20}	-1.441	-1.900	-0.981	0.842
α_{21}	1.550	1.552	1.228	1.876	0.942	α_{21}	1.373	0.837	1.910	0.868
α_{22}	0.750	0.752	0.642	0.861	0.930	α_{22}	0.321	0.012	0.631	0.270
σ	0.350	0.343	0.258	0.428	0.918	σ	0.930	0.808	1.052	0.000
Missingness mechanism										
β_0	3.450	3.491	2.806	4.176	0.946	β_0	6.446	3.451	9.441	0.478
β_1	-0.550	-0.556	-1.135	0.023	0.918	β_1	-3.458	-5.440	-1.475	0.098
β_2	-1.550	-1.571	-1.937	-1.205	0.960	β_2	-2.799	-3.961	-1.637	0.400
ψ	0.850	0.832	0.700	0.964	0.928	β_3	1.909	0.842	2.975	0.048
Nuisance parameter										
Markov Chain parameters										
π						π				
π_1	0.650	0.649				π_1	0.605			
π_2	0.350	0.351				π_2	0.395			
$\mathbb{Q}$						$\mathbb{Q}$				
q_{11}	0.400	0.377				q_{11}	0.781			
q_{12}	0.600	0.623				q_{12}	0.219			
q_{21}	0.350	0.370				q_{21}	0.060			
q_{22}	0.650	0.630				q_{22}	0.940			

[1] Simulation sample size $n = 150$. replicates $R = 500$.
[2] σ standard deviation of outcome at each time, ψ standard deviation of random effect b_i.
[3] $C.P$ coverage probability. [4] Rsp.Rate: response rate at each assessment time.
[5] 95% L: 95% lower confidence interval. [6] 95% U: 95% upper confidence interval.

Table 4.3: Simulation of model comparison $n = 300$: SHMM vs SPHMM

					n=300					
Rsp.Rate	$T_1 = 0.817$ $T_2 = 0.779$ $T_3 = 0.615$									
		SHMM					SPHMM			
		Correctly specified					Mis-specified			
	True	Est	95% L	95% U	C.P		Est	95% L	95% U	C.P
Parameters						Parameters				
α_{10}	0.650	0.651	0.584	0.719	0.950	α_{10}	0.643	0.577	0.709	0.938
α_{11}	1.050	1.054	0.939	1.169	0.948	α_{11}	0.961	0.861	1.060	0.590
α_{12}	0.250	0.248	0.177	0.319	0.948	α_{12}	0.219	0.158	0.281	0.830
α_{20}	-1.500	-1.497	-1.569	-1.425	0.950	α_{20}	-1.479	-1.550	-1.408	0.898
α_{21}	1.550	1.552	1.471	1.632	0.944	α_{21}	1.497	1.422	1.572	0.734
α_{22}	0.750	0.746	0.694	0.797	0.944	α_{22}	0.725	0.676	0.773	0.816
σ	0.350	0.348	0.310	0.385	0.940	σ	0.339	0.299	0.380	0.886
Missingness mechanism										
β_0	3.450	3.506	2.699	4.313	0.964	β_0	3.038	2.600	3.476	0.544
β_1	-0.550	-0.555	-1.387	0.278	0.972	β_1	-2.111	-2.492	-1.729	0.000
β_2	-0.350	-0.349	-0.596	-0.103	0.954	β_2	-0.641	-0.855	-0.427	0.258
β_3	-1.550	-1.570	-2.421	-0.719	0.980	ψ	-4.781	-7.911	-1.651	0.520
Nuisance parameter										
Markov Chain parameters										
π						π				
π_1	0.650	0.648				π_1	0.585			
π_2	0.350	0.352				π_2	0.415			
$\mathbb{Q}$						$\mathbb{Q}$				
q_{11}	0.400	0.379				q_{11}	0.305			
q_{12}	0.600	0.621				q_{12}	0.695			
q_{21}	0.350	0.374				q_{21}	0.296			
q_{22}	0.650	0.626				q_{22}	0.704			
					n=300					
Rsp.Rate	$T_1 = 0.945$ $T_2 = 0.803$ $T_3 = 0.515$									
		SPHMM					SHMM			
		Correctly specified					Mis-specified			
	True	Est	95% L	95% U	C.P		Est	95% L	95% U	C.P
Parameters						Parameters				
α_{10}	0.650	0.646	0.480	0.813	0.940	α_{10}	0.580	0.344	0.817	0.868
α_{11}	1.050	1.060	0.830	1.289	0.940	α_{11}	1.182	0.867	1.496	0.848
α_{12}	0.250	0.251	0.169	0.333	0.950	α_{12}	0.186	0.080	0.293	0.764
α_{20}	-1.500	-1.504	-1.685	-1.323	0.922	α_{20}	-1.422	-1.752	-1.093	0.844
α_{21}	1.550	1.559	1.328	1.790	0.950	α_{21}	1.371	0.993	1.750	0.840
α_{22}	0.750	0.750	0.672	0.828	0.952	α_{22}	0.300	0.088	0.512	0.030
σ	0.350	0.348	0.286	0.409	0.944	σ	0.953	0.869	1.036	0.000
Missingness mechanism										
β_0	3.450	3.476	2.994	3.958	0.960	β_0	6.080	4.285	7.876	0.122
β_1	-0.550	-0.553	-0.961	-0.144	0.948	β_1	-3.246	-4.445	-2.046	0.002
β_2	-1.550	-1.566	-1.824	-1.307	0.948	β_2	-2.657	-3.363	-1.951	0.068
ψ	0.850	0.845	0.751	0.939	0.954	β_3	1.783	1.152	2.414	0.000
Nuisance parameter										
Markov Chain parameters										
π						π				
π_1	0.650	0.650				π_1	0.603			
π_2	0.350	0.350				π_2	0.397			
$\mathbb{Q}$						$\mathbb{Q}$				
q_{11}	0.400	0.379				q_{11}	0.795			
q_{12}	0.600	0.621				q_{12}	0.205			
q_{21}	0.350	0.369				q_{21}	0.050			
q_{22}	0.650	0.631				q_{22}	0.950			

[1] Simulation sample size $n = 300$. replicates $R = 500$.
[2] σ standard deviation of outcome at each time, ψ standard deviation of random effect b_i.
[3] $C.P$ coverage probability. [4] Rsp.Rate: response rate at each assessment time.
[5] 95% L: 95% lower confidence interval. [6] 95% U: 95% upper confidence interval.

Table 4.4: Simulation of sensitivity analysis: transition matrix $\mathbb{Q}_A$

	n=500				
Rsp.Rate	$T_1 = 0.817$	$T_2 = 0.809$	$T_3 = 0.670$		
			SHMM		
	True	Est	95% L	95% U	C.P
Parameters					
α_{10}	0.650	0.655	0.600	0.709	0.958
α_{11}	1.050	1.035	0.943	1.128	0.942
α_{12}	0.250	0.204	0.105	0.304	0.824
α_{20}	-1.500	-1.491	-1.544	-1.437	0.938
α_{21}	1.550	1.545	1.488	1.602	0.952
α_{22}	0.750	0.735	0.700	0.770	0.850
σ	0.350	0.346	0.319	0.374	0.940
Missingness mechanism					
β_0	3.450	3.358	2.820	3.895	0.906
β_1	-0.550	-0.773	-1.496	-0.050	0.906
β_2	-0.350	-0.335	-0.511	-0.159	0.958
β_3	-1.550	-1.357	-2.016	-0.698	0.892
Nuisance parameter					
Markov Chain parameters					
π					
π_1	0.650	0.645			
π_2	0.350	0.355			
$\mathbb{Q}_{\mathbb{A}}$					
q_{11}	0.400	0.231			
q_{12}	0.600	0.769			
q_{21}	0.000	0.173			
q_{22}	1.000	0.827			

[1] Simulation sample size $n = 500$. replicates $R = 500$.
[2] σ standard deviation of outcome at each time.
[3] $C.P$ coverage probability. [4] Rsp.Rate: response rate at each assessment time. [5] 95% L: 95% lower confidence interval.
[6] 95% U: 95% upper confidence interval.

Table 4.5: Simulation of sensitivity analysis: transition matrix $\mathbb{Q}_B$

	n=500				
Rsp.Rate	$T_1 = 0.815$	$T_2 = 0.777$	$T_3 = 0.613$		
			SHMM		
	True	Est	95% L	95% U	C.P
Parameters					
α_{10}	0.650	0.655	0.585	0.725	0.962
α_{11}	1.050	1.037	0.895	1.178	0.956
α_{12}	0.250	0.179	0.007	0.350	0.872
α_{20}	-1.500	-1.417	-1.489	-1.345	0.386
α_{21}	1.550	1.487	1.405	1.569	0.674
α_{22}	0.750	0.878	0.825	0.931	0.000
σ	0.350	0.523	0.461	0.584	0.000
Missingness mechanism					
β_0	3.450	3.258	2.665	3.851	0.808
β_1	-0.550	-0.773	-1.427	-0.119	0.870
β_2	-0.350	-0.383	-0.568	-0.198	0.948
β_3	-1.550	-1.245	-1.900	-0.591	0.764
Nuisance parameter					
Markov Chain parameters					
π					
π_1	0.650	0.666			
π_2	0.350	0.334			
$\mathbb{Q}_{\mathbb{B}}$					
q_{11}	0.400	0.422			
q_{12}	0.600	0.578			
q_{21}	0.350	0.000*			
q_{22}	0.650	1.000*			

[1] Simulation sample size $n = 500$. replicates $R = 500$.
[2] σ standard deviation of outcome at each time.
[3] $C.P$ coverage probability. [4] Rsp.Rate: response rate at each assessment time. [5] 95% L: 95% lower confidence interval.
[6] 95% U: 95% upper confidence interval. [7] *: parameter are fixed at the number.

Table 4.6: Analysis for 5 years data: MMSE

	SPHMM				SHMM			
	State 1		State 2		State 1		State 2	
	Est	P-value	Est	P-value	Est	P-value	Est	P-value
Int	3.314	0.000	2.973	0.000	3.324	0.000	2.662	0.000
time	0.000	0.908	0.050	0.052	-0.002	0.274	-0.073	0.004
rx (PCV+RT)	0.000	0.964	0.698	0.000	0.003	0.699	0.405	0.000
resection (total resection)	-0.004	0.651	-0.568	0.000	0.000	0.957	0.040	0.677
kps (90-100)	0.047	0.000	-0.212	0.077	0.033	0.004	-0.781	0.000
grade (4-5)	0.007	0.477	-0.092	0.317	0.002	0.802	0.126	0.260
age (50+)	-0.038	0.002	-0.191	0.076	-0.032	0.006	-0.134	0.046
σ	0.065	0.000			0.090	0.000		
Missingness mechanism								
Int	1.910	0.000			-15.910	0.000		
time	-0.768	0.000			-0.689	0.000		
rx (PCV+RT)	-0.065	0.731			-1.073	0.057		
resection (total resection)	-0.137	0.474			0.127	0.782		
kps (90-100)	0.452	0.026			1.153	0.007		
grade (4-5)	-0.237	0.214			-0.248	0.506		
age (50+)	-0.586	0.003			-0.332	0.431		
ψ	0.071	0.000			Y_i 5.805	0.000		
Nuisance parameter								
Markov Chain parameters								
		π_1	π_2			π_1	π_2	
		0.972	0.028			0.969	0.031	
		$q_{.1}$	$q_{.2}$			$q_{.1}$	$q_{.2}$	
	$q_{1.}$	0.970	0.030		$q_{1.}$	0.848	0.152	
	$q_{2.}$	0.875	0.125		$q_{2.}$	0.029	0.971	

[1] σ standard deviation of outcome at each time.
[2] ψ standard deviation of random effect. [3] Est: Estimation [4] Category in parenthesis is of interest: baseline is first value of each variables in Table 4.1

Table 4.7: Analysis for 5 years data: B-QLQ

	SPHMM				SHMM			
	State 1		State 2		State 1		State 2	
	Est	P-value	Est	P-value	Est	P-value	Est	P-value
Int	4.129	0.000	4.147	0.000	4.238	0.000	3.952	0.000
time	-0.029	0.017	-0.010	0.622	-0.029	0.000	-0.275	0.000
rx (PCV+RT)	-0.044	0.355	0.052	0.310	0.027	0.427	-0.014	0.834
resection (total)	0.066	0.093	-0.101	0.070	-0.036	0.313	0.010	0.878
kps (90-100)	0.130	0.008	0.164	0.018	0.161	0.001	0.139	0.028
grade (4-5)	-0.045	0.412	0.043	0.539	-0.022	0.552	0.056	0.381
age (50+)	-0.008	0.822	-0.153	0.004	-0.126	0.010	-0.001	0.995
σ	0.173	0.000			0.228	0.000		
Missingness mechanism								
Int	1.753	0.000			-20.545	0.000		
time	-0.702	0.000			-0.368	0.000		
rx (PCV+RT)	-0.322	0.092			-0.257	0.385		
resection (total)	-0.326	0.094			-0.486	0.132		
kps (90-100)	0.532	0.009			-0.281	0.413		
grade (4-5)	-0.054	0.775			0.252	0.432		
age (50+)	-0.587	0.003			-0.496	0.129		
ψ	0.194	0.000			Y_i 5.671	0.000		
Nuisance parameter								
Markov Chain parameters								
		π_1	π_2			π_1	π_2	
		0.789	0.211			0.607	0.393	
		$q_{.1}$	$q_{.2}$			$q_{.1}$	$q_{.2}$	
	$q_{1.}$	0.501	0.499		$q_{1.}$	0.865	0.135	
	$q_{2.}$	0.489	0.511		$q_{2.}$	0.034	0.966	

[1] σ standard deviation of outcome at each time.
[2] ψ standard deviation of random effect. [3] Est: Estimation. [4] Category in parenthesis is of interest: baseline is first value of each variables in Table 4.1

Table 4.8: Patients characteristics by arm for at least 2 years survival

		PCV+RT		RT		
			%		%	p-value
		101	50.25	100	49.75	
Age						0.553
	50 >	72	71.29	75	75.00	
	50 <	29	28.71	25	25.00	
Resection						0.1387
	biopsy/partial	41	40.59	51	51.00	
	total resection	60	59.41	49	49.00	
KPS						0.239
	60–80	22	21.78	29	29.00	
	90–100	79	78.22	71	71.00	
Grade						0.704
	anaplastic (2-3 features)	61	60.40	63	63.00	
	anaplastic (4-5 features)	40	39.60	37	37.00	

Table 4.9: Data analysis for at least 2 years survival: MMSE

	SPHMM				SHMM			
	State 1		State 2		State 1		State 2	
	Est	P-value	Est	P-value	Est	P-value	Est	P-value
Int	3.322	0.000	2.647	0.000	3.321	0.000	2.639	0.000
time	0.007	0.061	0.234	0.000	0.006	0.066	0.219	0.000
rx (PCV+RT)	0.003	0.708	0.102	0.079	0.004	0.552	0.113	0.031
resection (total)	-0.016	0.045	-0.127	0.013	-0.013	0.062	-0.157	0.000
kps (90-100)	0.044	0.000	-0.077	0.157	0.042	0.000	-0.053	0.228
grade (4-5)	0.012	0.158	0.309	0.000	0.013	0.077	0.261	0.000
age (50+)	-0.024	0.011	-0.200	0.011	-0.025	0.008	-0.145	0.008
σ	0.063	0.000			0.069	0.000		
Missingness mechanism								
Int	2.804	0.000			-11.549	0.005		
time	-1.317	0.000			-1.410	0.000		
rx (PCV+RT)	-0.145	0.606			-0.281	0.377		
resection (total)	-0.038	0.893			0.241	0.468		
kps (90-100)	0.436	0.133			0.206	0.525		
grade (4-5)	0.391	0.167			0.058	0.867		
age (50+)	-0.064	0.814			0.223	0.473		
ψ	0.029	0.000			Y_i 4.459	0.001		
Nuisance parameter								
Markov Chain parameters								
		π_1	π_2			π_1	π_2	
		0.984	0.016			0.985	0.015	
		$q_{.1}$	$q_{.2}$			$q_{.1}$	$q_{.2}$	
	$q_{1.}$	0.946	0.054		$q_{1.}$	0.927	0.073	
	$q_{2.}$	0.350	0.650		$q_{2.}$	0.003	0.997	

[1] σ standard deviation of outcome at each time.
[2] ψ standard deviation of random effect. [3] Est: Estimation. [4] Category in parenthesis is of interest: baseline is first value of each variables in Table 4.1

Table 4.10: Data analysis for at least 2 years survival: B-QOL

	SPHMM				SHMM			
	State 1		State 2		State 1		State 2	
	Est	P-value	Est	P-value	Est	P-value	Est	P-value
Int	4.167	0.000	3.853	0.000	4.143	0.000	3.830	0.000
time	0.020	0.106	-0.177	0.008	-0.014	0.481	-0.001	0.970
rx (PCV+RT)	-0.043	0.215	-0.240	0.021	-0.069	0.135	-0.042	0.338
resection (total)	-0.028	0.440	0.229	0.031	-0.004	0.925	0.007	0.828
kps (90-100)	0.116	0.002	0.734	0.000	0.096	0.051	0.606	0.000
grade (4-5)	0.033	0.324	-0.268	0.001	0.136	0.006	-0.425	0.000
age (50+)	-0.030	0.442	-0.116	0.233	-0.055	0.307	0.034	0.466
σ	0.121	0.000			0.069	0.000		
Missingness mechanism								
Int	2.043	0.000			-11.549	0.005		
time	-0.811	0.000			-1.410	0.000		
rx (PCV+RT)	-0.379	0.122			-0.281	0.377		
resection (total)	-0.280	0.271			0.241	0.468		
kps (90-100)	0.385	0.163			0.206	0.525		
grade (4-5)	0.550	0.033			0.058	0.867		
age (50+)	-0.147	0.570			0.223	0.473		
ψ	0.188	0.000			Y_i 4.459	0.001		
Nuisance parameter								
Markov Chain parameters								
		π_1	π_2			π_1	π_2	
		0.875	0.125			0.771	0.229	
		$q_{.1}$	$q_{.2}$			$q_{.1}$	$q_{.2}$	
	$q_{1.}$	0.831	0.169		$q_{1.}$	0.847	0.153	
	$q_{2.}$	0.711	0.289		$q_{2.}$	0.062	0.938	

[1] σ standard deviation of outcome at each time.
[2] ψ standard deviation of random effect. [3] Est: Estimation. [4] Category in parenthesis is of interest: baseline is first value of each variables in Table 4.1

CHAPTER 5 : Conclusion

In this dissertation, we have developed new statistical methods to handle non-monotone and non-ignorable missing data in longitudinal studies. We assume a first-order Markov structure in both the complete data and missingness mechanism, which is a natural way to capture the changes in outcomes among repeated measurements in a longitudinal data setting and to properly accommodate the variance-covariance structure. In Chapter 2, we developed a full-likelihood method to analyze continuous longitudinal responses with non-ignorable non-monotone missing data. This method is an extension of the work of Troxel et al. (1998a). We adopt the multivariate Gaussian distribution assumption for the underlying data and a first-order Markov dependence structure. Instead of using logistic regression to model the missing mechanism, we propose a beta-binomial distribution to model the probability of non-response. The multivariate Polya distribution is a high-dimensional version of the beta-binomial distribution; the beta and binomial distributions correspond to Dirichlet and multinomial distributions, respectively, in the multivariate situation. This helps to stabilize the estimation of the missingness mechanism, especially when some time points have small amounts of missing or no missing data. This mixture model also reduces multimodality in the likelihood. This method has better power and more robust performance for parameter estimation. We conducted simulations to demonstrate the empirical behavior of the proposed models as well. In Chapter 3, we developed a transition pseudo-likelihood approach by considering only adjacent pairs of observations. This method can be viewed as an extension of composite marginal likelihood methods (Cox and Reid, 2004; Varin et al., 2011) with application to the non-ignorable non-monotone missing data framework. This pseudo-likelihood approach can significantly reduce the computational complexities of the full-likelihood based method. The transition pseudo-score function is used to obtain correct inference in spite of the independence assumption among the sets of adjacent pairs. The simulation study shows that this approach can handle longitudinal data with various covariance structures well and is no more computationally intensive than the

independent pseudo-likelihood model (Troxel et al., 1998b), which makes this model attractive for situations with a large number of assessments. In Chapter 4, we further consider a hidden Markov model incorporating both selection models and shared parameter models to capture the disease progression through different hidden states. The conditional independence assumed in the hidden Markov model provides a simple framework for reducing the multi-dimensional integration in traditional methods into one-dimensional integration in the observed likelihood. In addition, the proposed models avoid the problem of specification of the correlation structure of repeated outcomes instead of emphasizing estimation in Markov Chain parameters. A two stage pseudo-likelihood algorithm was used to reduce the parameter space and obtain inference. This approach allows more precise identification of the marginal effects. Simulation studies were conducted to investigate the empirical behavior of the proposed models. Sensitivity analyses were provided to evaluate the method's performance when Markov Chain parameters are mis-specified.

In summary, we have developed several novel statistical methods for handling non-monotone and non-ignorable missing data in longitudinal studies. Model selection differs depending on the outcome process and the missingness mechanism. Derived from the theory of pseudo likelihood-based methods, the proposed pseudo likelihood-based approach requires a large sample size to improve the performance. As with any model-based approach to non-ignorable missing data, the current approach is subject to unavoidable assumptions about the complete data distribution and the missing data mechanism. It is important to consider all substantive information about the area of application, prior experience with missing data in similar situations, and expert opinion about the mechanism of missing data when building such models. In many areas, enough knowledge and experience exists to justify the necessary assumptions, and the benefit in terms of bias reduction can be significant.

APPENDIX

A.1. Conditional Density

For $T = 3$, assume the first observation does depend other covariate and is always observed.

For T=1 then

$$\begin{aligned} f(y_{i1}) &= \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(\frac{1}{2\sigma_1^2}(y_{i1} - \mu_{i1})^2) \\ f(R_{i1}|y_{i1}) &= \pi_{i1}^{R_{i1}}(1 - \pi_{i1})^{1-R_{i1}} \end{aligned}$$

For T=2 then

$$\begin{aligned} f(y_{i2}|y_{i1}) &= \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho_1^2)}} \exp(\frac{1}{2\sigma_2^2(1-\rho_1^2)}(y_{i2} - \mu_{i2} - \rho_1\frac{\sigma_2}{\sigma_1}(y_{i1} - \mu_{i1}))^2) \\ f(R_{i2} = 1|R_{i1}, y_{i2}) &= \frac{\exp(\beta_1^{'}\mathbf{X_{i2}} + \theta_\mathbf{1}\mathbf{Y_{i2}} + \psi_\mathbf{1}\mathbf{R_{i1}})}{\sum_{s=1}^{2}\exp(\beta_s^{'}\mathbf{X_{i2}} + \theta_\mathbf{s}\mathbf{Y_{i2}} + \psi_\mathbf{s}\mathbf{R_{i1}})}. \end{aligned}$$

For T=3 then

$$\begin{aligned} f(y_{i3}|y_{i2}) &= \frac{1}{\sqrt{2\pi\sigma_3^2(1-\rho_2^2)}} \exp(\frac{1}{2\sigma_3^2(1-\rho_2^2)}(y_{i3} - \mu_{i3} - \rho_2\frac{\sigma_3}{\sigma_2}(y_{i2} - \mu_{i2}))^2) \\ f(R_{i3} = 1|R_{i2}, y_{i3}) &= \frac{\exp(\beta_1^{'}\mathbf{X_{i3}} + \theta_\mathbf{1}\mathbf{Y_{i3}} + \psi_\mathbf{1}\mathbf{R_{i2}})}{\sum_{s=1}^{2}\exp(\beta_s^{'}\mathbf{X_{i3}} + \theta_\mathbf{s}\mathbf{Y_{i3}} + \psi_\mathbf{s}\mathbf{R_{i2}})}. \end{aligned}$$

A.2. Joint Likelihood Function

For $T = 3$ there are $2^3 = 8$ patterns, if don't allow all points to be missing $\begin{pmatrix} * & * & * \\ R_{i1} & R_{i2} & R_{i3} \\ 0 & 0 & 0 \end{pmatrix}$.

then we will have 7 patterns. I will list all possible patterns below.

Pattern 1 $P_1 := \begin{pmatrix} Y_{i1} & Y_{i2} & Y_{i3} \\ R_{i1} & R_{i2} & R_{i3} \\ 1 & 1 & 1 \end{pmatrix}$.

$$
\begin{aligned}
& \mathcal{L}^{p1}_{i,obs} = \mathcal{L}_i \\
= \quad & f(y_{i1})f(y_{i2}|y_{i1})f(y_{i3}|y_{i2})f(R_{i1}|y_{i1})f(R_{i2}|R_{i1},y_{i2})f(R_{i3}|R_{i2},y_{i3}) \\
= \quad & \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp(\frac{-1}{2\sigma_1^2}(y_{i1}-\mu_{i1})^2)\pi_{i1} \\
\times \quad & \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho_1^2)}}\exp(\frac{-1}{2\sigma_2^2(1-\rho_1^2)}(y_{i2}-\mu_{i2}-\rho_1\frac{\sigma_2}{\sigma_1}(y_{i1}-\mu_{i1}))^2)\frac{\exp(\beta_1^{'}\mathbf{X_{i2}}+\theta_1\mathbf{Y_{i2}}+\psi_1)}{\sum_{s=1}^{2}\exp(\beta_s^{'}\mathbf{X_{i2}}+\theta_s\mathbf{Y_{i2}}+\psi_s)} \\
\times \quad & \frac{1}{\sqrt{2\pi\sigma_3^2(1-\rho_2^2)}}\exp(\frac{-1}{2\sigma_3^2(1-\rho_2^2)}(y_{i3}-\mu_{i3}-\rho_2\frac{\sigma_3}{\sigma_2}(y_{i2}-\mu_{i2}))^2)\frac{\exp(\beta_1^{'}\mathbf{X_{i3}}+\theta_1\mathbf{Y_{i3}}+\psi_1)}{\sum_{s=1}^{2}\exp(\beta_s^{'}\mathbf{X_{i3}}+\theta_s\mathbf{Y_{i3}}+\psi_s)}
\end{aligned}
$$

Pattern 2 $P_2 := \begin{pmatrix} Y_{i1} & Y_{i2} & * \\ R_{i1} & R_{i2} & R_{i3} \\ 1 & 1 & 0 \end{pmatrix}$.

$$
\begin{aligned}
& \mathcal{L}^{p2}_{i,obs} = \int \mathcal{L}_i dy_{i3} \\
= \quad & \int f(y_{i1})f(y_{i2}|y_{i1})f(y_{i3}|y_{i2})f(R_{i1}|y_{i1})f(R_{i2}|R_{i1},y_{i2})f(R_{i3}|R_{i2},y_{i3})dy_{i3} \\
= \quad & \int f(R_{i3}|R_{i2},y_{i3})f(y_{i3}|y_{i2})dy_{i3} \\
\times \quad & f(y_{i1})f(y_{i2}|y_{i1})f(R_{i1}|y_{i1})f(R_{i2}|R_{i1},y_{i2}) \\
= \quad & \mathbb{E}_{f_{3|2}}(f(R_{i3}|R_{i2},y_{i3}^{*}))\times f(y_{i1})f(y_{i2}|y_{i1})f(R_{i1}|y_{i1})f(R_{i2}|R_{i1},y_{i2}) \\
= \quad & \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp(\frac{-1}{2\sigma_1^2}(y_{i1}-\mu_{i1})^2)\pi_{i1} \\
\times \quad & \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho_1^2)}}\exp(\frac{-1}{2\sigma_2^2(1-\rho_1^2)}(y_{i2}-\mu_{i2}-\rho_1\frac{\sigma_2}{\sigma_1}(y_{i1}-\mu_{i1}))^2)\frac{\exp(\beta_1^{'}\mathbf{X_{i2}}+\theta_1\mathbf{Y_{i2}}+\psi_1)}{\sum_{s=1}^{2}\exp(\beta_s^{'}\mathbf{X_{i2}}+\theta_s\mathbf{Y_{i2}}+\psi_s)} \\
\times \quad & \int \frac{1}{\sqrt{2\pi\sigma_3^2(1-\rho_2^2)}}\exp(\frac{-1}{2\sigma_3^2(1-\rho_2^2)}(y_{i3}-\mu_{i3}-\rho_2\frac{\sigma_3}{\sigma_2}(y_{i2}-\mu_{i2}))^2)\frac{\exp(\beta_2^{'}\mathbf{X_{i3}}+\theta_2\mathbf{Y_{i3}}+\psi_2)}{\sum_{s=1}^{2}\exp(\beta_s^{'}\mathbf{X_{i3}}+\theta_s\mathbf{Y_{i3}}+\psi_s)}dy_{i3} \\
= \quad & \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp(\frac{-1}{2\sigma_{i1}^2}(y_{i1}-\mu_{i1})^2)\pi_{i1} \\
\times \quad & \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho_1^2)}}\exp(\frac{-1}{2\sigma_2^2(1-\rho_1^2)}(y_{i2}-\mu_{i2}-\rho_1\frac{\sigma_2}{\sigma_1}(y_{i1}-\mu_{i1}))^2)\frac{\exp(\beta_1^{'}\mathbf{X_{i2}}+\theta_1\mathbf{Y_{i2}}+\psi_1)}{\sum_{s=1}^{2}\exp(\beta_s^{'}\mathbf{X_{i2}}+\theta_s\mathbf{Y_{i2}}+\psi_s)} \\
\times \quad & \sum_{k=1}^{m}\frac{w_k}{\sqrt{\pi}}\frac{\exp(\beta_2^{'}\mathbf{X_{i3}}+\theta_2(\mu_{i3}+\rho_2\frac{\sigma_3}{\sigma_2}(\mathbf{y_{i2}}-\mu_{i2})+\sqrt{2\sigma_3^2(1-\rho_2^2)}\tau_k)+\psi_2)}{\sum_{s=1}^{2}\exp(\beta_s^{'}\mathbf{X_{i3}}+\theta_s(\mu_{i3}+\rho_2\frac{\sigma_3}{\sigma_2}(\mathbf{y_{i2}}-\mu_{i2})+\sqrt{2\sigma_3^2(1-\rho_2^2)}\tau_k)+\psi_s)}
\end{aligned}
$$

Pattern 3 $P_3 := \begin{pmatrix} Y_{i1} & * & * \\ R_{i1} & R_{i2} & R_{i3} \\ 1 & 0 & 0 \end{pmatrix}$

$$
\begin{aligned}
& \mathcal{L}^{p3}_{i,obs} = \int\int \mathcal{L}_i dy_{i2} dy_{i3} \\
= \;& \int\int f(R_{i3}|R_{i2}, y_{i3}) f(R_{i2}|R_{i1}, y_{i2}) f(y_{i3}|y_{i2}) f(y_{i2}|y_{i1}) dy_{i2} dy_{i3} f(R_{i1}|y_{i1}) f(y_{i1}) \\
= \;& \int\int f(R_{i3}|R_{i2}, y_{i3}) f(R_{i2}|R_{i1}, y_{i2}) f(y_{i3}, y_{i2}|y_{i1}) dy_{i2} dy_{i3} f(R_{i1}|y_{i1}) f(y_{i1}) \\
= \;& \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(\frac{-1}{2\sigma_1^2}(y_{i1} - \mu_{i1})^2) \pi_{i1} \\
\times \;& \int\int \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho_1^2)}} \exp(\frac{-1}{2\sigma_2^2(1-\rho_1^2)}(y_{i2} - \mu_{i2} - \rho_1 \frac{\sigma_2}{\sigma_1}(y_{i1} - \mu_{i1}))^2) \frac{\exp(\beta_2' \mathbf{X_{i2}} + \theta_2 \mathbf{Y_{i2}} + \psi_2)}{\sum_{s=1}^2 \exp(\beta_s' \mathbf{X_{i2}} + \theta_s \mathbf{Y_{i2}} + \psi_s)} \\
\times \;& \frac{1}{\sqrt{2\pi\sigma_3^2(1-\rho_2^2)}} \exp(\frac{-1}{2\sigma_3^2(1-\rho_2^2)}(y_{i3} - \mu_{i3} - \rho_2 \frac{\sigma_3}{\sigma_2}(y_{i2} - \mu_{i2}))^2) \frac{\exp(\beta_2' \mathbf{X_{i3}} + \theta_2 \mathbf{Y_{i3}})}{\sum_{s=1}^2 \exp(\beta_s' \mathbf{X_{i3}} + \theta_s \mathbf{Y_{i3}})} dy_{i2} dy_{i3} \\
= \;& \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(\frac{-1}{2\sigma_{i1}^2}(y_{i1} - \mu_{i1})^2) \pi_{i1} \\
\times \;& \int \sum_{k=1}^m \frac{w_k}{\sqrt{\pi}} \frac{\exp(\beta_2' \mathbf{X_{i2}} + \theta_2(\mu_{i2} + \rho_1 \frac{\sigma_2}{\sigma_1}(\mathbf{y_{i1}} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k) + \psi_2)}{\sum_{s=1}^2 \exp(\beta_s' \mathbf{X_{i2}} + \theta_s(\mu_{i2} + \rho_1 \frac{\sigma_2}{\sigma_1}(\mathbf{y_{i1}} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k) + \psi_s)} \\
\times \;& \frac{1}{\sqrt{2\pi\sigma_3^2(1-\rho_2^2)}} \exp(\frac{-1}{2\sigma_3^2(1-\rho_2^2)}(y_{i3} - \mu_{i3} - \rho_2 \frac{\sigma_3}{\sigma_2}((\mu_{i2} + \rho_1 \frac{\sigma_2}{\sigma_1}(y_{i1} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k) - \mu_{i2}))^2) \\
\times \;& \frac{\exp(\beta_2' \mathbf{X_{i3}} + \theta_2 \mathbf{Y_{i3}})}{\sum_{s=1}^2 \exp(\beta_s' \mathbf{X_{i3}} + \theta_s \mathbf{Y_{i3}})} dy_{i3} \\
= \;& \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(\frac{-1}{2\sigma_{i1}^2}(y_{i1} - \mu_{i1})^2) \pi_{i1} \\
\times \;& \sum_{k=1}^m \frac{w_k}{\sqrt{\pi}} \frac{\exp(\beta_2' \mathbf{X_{i2}} + \theta_2(\mu_{i2} + \rho_1 \frac{\sigma_2}{\sigma_1}(\mathbf{y_{i1}} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k) + \psi_2)}{\sum_{s=1}^2 \exp(\beta_s' \mathbf{X_{i2}} + \theta_s(\mu_{i2} + \rho_1 \frac{\sigma_2}{\sigma_1}(\mathbf{y_{i1}} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k) + \psi_s)} \\
\times \;& \int \frac{1}{\sqrt{2\pi\sigma_3^2(1-\rho_2^2)}} \exp(\frac{-1}{2\sigma_3^2(1-\rho_2^2)}(y_{i3} - \mu_{i3} - \rho_2 \frac{\sigma_3}{\sigma_2}(\rho_1 \frac{\sigma_2}{\sigma_1}(y_{i1} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k))^2) \\
\times \;& \frac{\exp(\beta_2' \mathbf{X_{i3}} + \theta_2 \mathbf{Y_{i3}})}{\sum_{s=1}^2 \exp(\beta_s' \mathbf{X_{i3}} + \theta_s \mathbf{Y_{i3}})} dy_{i3} \\
= \;& \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(\frac{-1}{2\sigma_1^2}(y_{i1} - \mu_{i1})^2) \pi_{i1} \\
\times \;& \sum_{k=1}^m \frac{w_k}{\sqrt{\pi}} \frac{\exp(\beta_2' \mathbf{X_{i2}} + \theta_2(\mu_{i2} + \rho_1 \frac{\sigma_2}{\sigma_1}(\mathbf{y_{i1}} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k) + \psi_2)}{\sum_{s=1}^2 \exp(\beta_s' \mathbf{X_{i2}} + \theta_s(\mu_{i2} + \rho_1 \frac{\sigma_2}{\sigma_1}(\mathbf{y_{i1}} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k) + \psi_s)} \\
\times \;& \sum_{l=1}^m \frac{w_l}{\sqrt{\pi}} \frac{\exp(\beta_2' \mathbf{X_{i3}} + \theta_2(\mu_{i3} + \rho_2 \frac{\sigma_3}{\sigma_2}(\rho_1 \frac{\sigma_2}{\sigma_1}(\mathbf{y_{i1}} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k) + \sqrt{2\pi\sigma_3^2(1-\rho_2^2)}\tau_l))}{\sum_{s=1}^2 \exp(\beta_s' \mathbf{X_{i3}} + \theta_s(\mu_{i3} + \rho_2 \frac{\sigma_3}{\sigma_2}(\rho_1 \frac{\sigma_2}{\sigma_1}(\mathbf{y_{i1}} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k) + \sqrt{2\pi\sigma_3^2(1-\rho_2^2)}\tau_l))} \\
= \;& \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(\frac{-1}{2\sigma_1^2}(y_{i1} - \mu_{i1})^2) \pi_{i1} \\
\times \;& \sum_{k=1}^m \sum_{l=1}^m \frac{w_k w_l}{\pi} \frac{\exp(\beta_2' \mathbf{X_{i2}} + \theta_2(\mu_{i2} + \rho_1 \frac{\sigma_2}{\sigma_1}(\mathbf{y_{i1}} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k) + \psi_2)}{\sum_{s=1}^2 \exp(\beta_s' \mathbf{X_{i2}} + \theta_s(\mu_{i2} + \rho_1 \frac{\sigma_2}{\sigma_1}(\mathbf{y_{i1}} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k) + \psi_s)} \\
\times \;& \frac{\exp(\beta_2' \mathbf{X_{i3}} + \theta_2(\mu_{i3} + \rho_2 \frac{\sigma_3}{\sigma_2}(\rho_1 \frac{\sigma_2}{\sigma_1}(\mathbf{y_{i1}} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k) + \sqrt{2\pi\sigma_3^2(1-\rho_2^2)}\tau_l))}{\sum_{s=1}^2 \exp(\beta_s' \mathbf{X_{i3}} + \theta_s(\mu_{i3} + \rho_2 \frac{\sigma_3}{\sigma_2}(\rho_1 \frac{\sigma_2}{\sigma_1}(\mathbf{y_{i1}} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k) + \sqrt{2\pi\sigma_3^2(1-\rho_2^2)}\tau_l))}
\end{aligned}
$$

Pattern 4 $P_4 := \begin{pmatrix} Y_{i1} & * & Y_{i3} \\ R_{i1} & R_{i2} & R_{i3} \\ 1 & 0 & 1 \end{pmatrix}$.

$$
\begin{aligned}
& \mathcal{L}^{p4}_{i,obs} = \int \mathcal{L}_i dy_{i2} \\
= & \int f(y_{i1}) f(y_{i2}|y_{i1}) f(y_{i3}|y_{i2}) f(R_{i1}|y_{i1}) f(R_{i2}|R_{i1}, y_{i2}) f(R_{i3}|R_{i2}, y_{i3}) dy_{i2} \\
= & \int f(R_{i2}|R_{i1}, y_{i2}) f(y_{i3}|y_{i2}) f(y_{i2}|y_{i1}) dy_{i2} \\
\times & f(y_{i1}) f(R_{i1}|y_{i1}) f(R_{i3}|R_{i2}, y_{i3}) \\
= & \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(\frac{-1}{2\sigma_1^2}(y_{i1} - \mu_{i1})^2) \pi_{i1} \frac{\exp(\beta_1^{'} \mathbf{X_{i3}} + \theta_1 \mathbf{Y_{i3}})}{\sum_{s=1}^{2} \exp(\beta_s^{'} \mathbf{X_{i3}} + \theta_s \mathbf{Y_{i3}})} \\
\times & \int \frac{\exp(\beta_2^{'} \mathbf{X_{i2}} + \theta_2 \mathbf{Y_{i2}} + \psi_2)}{\sum_{s=1}^{2} \exp(\beta_s^{'} \mathbf{X_{i2}} + \theta_s \mathbf{Y_{i2}} + \psi_s)} \frac{1}{\sqrt{2\pi\sigma_3^2(1-\rho_2^2)}} \exp(\frac{-1}{2\sigma_3^2(1-\rho_2^2)}(y_{i3} - \mu_{i3} - \rho_2 \frac{\sigma_3}{\sigma_2}(y_{i2} - \mu_{i2}))^2) \\
\times & \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho_1^2)}} \exp(\frac{-1}{2\sigma_2^2(1-\rho_1^2)}(y_{i2} - \mu_{i2} - \rho_1 \frac{\sigma_2}{\sigma_1}(y_{i1} - \mu_{i1}))^2) dy_{i2} \\
= & \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(\frac{-1}{2\sigma_1^2}(y_{i1} - \mu_{i1})^2) \pi_{i1} \frac{\exp(\beta_1^{'} \mathbf{X_{i3}} + \theta_1 \mathbf{Y_{i3}})}{\sum_{s=1}^{2} \exp(\beta_s^{'} \mathbf{X_{i3}} + \theta_s \mathbf{Y_{i3}})} \\
\times & \sum_{k=1}^{m} \frac{w_k}{\sqrt{\pi}} \frac{\exp(\beta_2^{'} \mathbf{X_{i2}} + \theta_2 (\mu_{i2} + \rho_1 \frac{\sigma_2}{\sigma_1}(\mathbf{y_{i1}} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k) + \psi_2)}{\sum_{s=1}^{2} \exp(\beta_s^{'} \mathbf{X_{i2}} + \theta_s (\mu_{i2} + \rho_1 \frac{\sigma_2}{\sigma_1}(\mathbf{y_{i1}} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k) + \psi_s)} \\
\times & \frac{1}{\sqrt{2\pi\sigma_3^2(1-\rho_2^2)}} \exp(\frac{-1}{2\sigma_3^2(1-\rho_2^2)}(y_{i3} - \mu_{i3} - \rho_2 \frac{\sigma_3}{\sigma_2}((\mu_{i2} + \rho_1 \frac{\sigma_2}{\sigma_1}(y_{i1} - \mu_{i1}) + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_k) - \mu_{i2}))^2)
\end{aligned}
$$

Pattern 5 $P_5 := \begin{pmatrix} * & Y_{i2} & Y_{i3} \\ R_{i1} & R_{i2} & R_{i3} \\ 0 & 1 & 1 \end{pmatrix}$

$$
\begin{aligned}
& \mathcal{L}^{p5}_{i,obs} = \int \mathcal{L}_i dy_{i1} \\
= & \int f(R_{i1}|y_{i1}) f(y_{i2}|y_{i1}) f(y_{i1}) dy_{i1} f(R_{i3}|R_{i2}, y_{i3}) f(R_{i2}|R_{i1}, y_{i2}) f(y_{i3}|y_{i2}) \\
= & \frac{\exp(\beta_1^{'} \mathbf{X_{i3}} + \theta_1 \mathbf{Y_{i3}} + \psi_1)}{\sum_{s=1}^{2} \exp(\beta_s^{'} \mathbf{X_{i3}} + \theta_s \mathbf{Y_{i3}} + \psi_s)} \frac{\exp(\beta_1^{'} \mathbf{X_{i2}} + \theta_1 \mathbf{Y_{i2}})}{\sum_{s=1}^{2} \exp(\beta_s^{'} \mathbf{X_{i2}} + \theta_s \mathbf{Y_{i2}})} \\
\times & \frac{1}{\sqrt{2\pi\sigma_3^2(1-\rho_2^2)}} \exp(\frac{-1}{2\sigma_3^2(1-\rho_2^2)}(y_{i3} - \mu_{i3} - \rho_2 \frac{\sigma_3}{\sigma_2}(y_{i2} - \mu_{i2}))^2) \\
\times & (1 - \pi_{i1}) \int \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho_1^2)}} \exp(\frac{-1}{2\sigma_2^2(1-\rho_1^2)}(y_{i2} - \mu_{i2} - \rho_1 \frac{\sigma_2}{\sigma_1}(y_{i1} - \mu_{i1}))^2) \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(\frac{1}{2\sigma_1^2}(y_{i1} - \mu_{i1})^2) dy_{i1} \\
= & (1 - \pi_{i1}) \frac{\exp(\beta_1^{'} \mathbf{X_{i3}} + \theta_1 \mathbf{Y_{i3}} + \psi_1)}{\sum_{s=1}^{2} \exp(\beta_s^{'} \mathbf{X_{i3}} + \theta_s \mathbf{Y_{i3}} + \psi_s)} \frac{\exp(\beta_1^{'} \mathbf{X_{i2}} + \theta_1 \mathbf{Y_{i2}})}{\sum_{s=1}^{2} \exp(\beta_s^{'} \mathbf{X_{i2}} + \theta_s \mathbf{Y_{i2}})} \\
\times & \frac{1}{\sqrt{2\pi\sigma_3^2(1-\rho_2^2)}} \exp(\frac{-1}{2\sigma_3^2(1-\rho_2^2)}(y_{i3} - \mu_{i3} - \rho_2 \frac{\sigma_3}{\sigma_2}(y_{i2} - \mu_{i2}))^2) \\
\times & \sum_{k=1}^{m} \frac{w_k}{\sqrt{\pi}} \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho_1^2)}} \exp(\frac{-1}{2\sigma_2^2(1-\rho_1^2)}(y_{i2} - \mu_{i2} - \rho_1 \frac{\sigma_2}{\sigma_1}((\mu_{i1} + \sqrt{2\sigma_1^2}\tau_k) - \mu_{i1}))^2)
\end{aligned}
$$

Pattern 6 $P_6 := \begin{pmatrix} * & Y_{i2} & * \\ R_{i1} & R_{i2} & R_{i3} \\ 0 & 1 & 0 \end{pmatrix}$

$$
\begin{aligned}
& \mathcal{L}^{p6}_{i,obs} = \int\int \mathcal{L}_i dy_{i1} dy_{i3} \\
= \;& \int\int f(y_{i1}) f(y_{i2}|y_{i1}) f(y_{i3}|y_{i2}) f(R_{i1}|y_{i1}) f(R_{i3}|R_{i2}, y_{i3}) dy_{i1} dy_{i3} f(R_{i2}|R_{i1}, y_{i2}) \\
= \;& \int f(R_{i3}|R_{i2}, y_{i3}) f(y_{i3}|y_{i2}) dy_{i3} \times \int f(y_{i1}) f(y_{i2}|y_{i1}) f(R_{i1}|y_{i1}) dy_{i1} \times f(R_{i2}|R_{i1}, y_{i2}) \\
= \;& (1-\pi_{i1}) \frac{\exp(\beta_1^{'} \mathbf{X_{i2}} + \boldsymbol{\theta_1} \mathbf{Y_{i2}})}{\sum_{s=1}^{2} \exp(\beta_s^{'} \mathbf{X_{i2}} + \boldsymbol{\theta_s} \mathbf{Y_{i2}})} \\
\times \;& \int \frac{\exp(\beta_2^{'} \mathbf{X_{i3}} + \boldsymbol{\theta_2} \mathbf{Y_{i3}} + \boldsymbol{\psi_2})}{\sum_{s=1}^{2} \exp(\beta_s^{'} \mathbf{X_{i3}} + \boldsymbol{\theta_s} \mathbf{Y_{i3}} + \boldsymbol{\psi_s})} \frac{1}{\sqrt{2\pi\sigma_3^2(1-\rho_2^2)}} \exp(\frac{-1}{2\sigma_3^2(1-\rho_2^2)} (y_{i3} - \mu_{i3} - \rho_2 \frac{\sigma_3}{\sigma_2}(y_{i2} - \mu_{i2}))^2) dy_{i3} \\
\times \;& \int \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho_1^2)}} \exp(\frac{-1}{2\sigma_2^2(1-\rho_1^2)} (y_{i2} - \mu_{i2} - \rho_1 \frac{\sigma_2}{\sigma_1}(y_{i1} - \mu_{i1}))^2) \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(\frac{1}{2\sigma_1^2}(y_{i1} - \mu_{i1})^2) dy_{i1} \\
= \;& (1-\pi_{i1}) \frac{\exp(\beta_1^{'} \mathbf{X_{i2}} + \boldsymbol{\theta_1} \mathbf{Y_{i2}})}{\sum_{s=1}^{2} \exp(\beta_s^{'} \mathbf{X_{i2}} + \boldsymbol{\theta_s} \mathbf{Y_{i2}})} \\
\times \;& \sum_{k=1}^{m} \sum_{l=1}^{m} \frac{w_k w_l}{\pi} \frac{\exp(\beta_2^{'} \mathbf{X_{i3}} + \boldsymbol{\theta_2}(\mu_{\mathbf{i3}} + \rho_{\mathbf{2}} \frac{\sigma_{\mathbf{3}}}{\sigma_{\mathbf{2}}}(\mathbf{y_{i2}} - \mu_{\mathbf{i2}}) + \sqrt{\mathbf{2}\sigma_{\mathbf{3}}^{\mathbf{2}}(\mathbf{1}-\rho_{\mathbf{2}}^{\mathbf{2}})}\tau_{\mathbf{k}}) + \boldsymbol{\psi_2})}{\sum_{s=1}^{2} \exp(\beta_s^{'} \mathbf{X_{i3}} + \boldsymbol{\theta_s} * (\mu_{\mathbf{i3}} + \rho_{\mathbf{2}} \frac{\sigma_{\mathbf{3}}}{\sigma_{\mathbf{2}}}(\mathbf{y_{i2}} - \mu_{\mathbf{i2}}) + \sqrt{\mathbf{2}\sigma_{\mathbf{3}}^{\mathbf{2}}(\mathbf{1}-\rho_{\mathbf{2}}^{\mathbf{2}})}\tau_{\mathbf{k}}) + \boldsymbol{\psi_s})} \\
\times \;& \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho_1^2)}} \exp(\frac{-1}{2\sigma_2^2(1-\rho_1^2)} (y_{i2} - \mu_{i2} - \sqrt{2}\rho_1\sigma_2\tau_l)^2)
\end{aligned}
$$

Pattern 7 $P_7 := \begin{pmatrix} * & * & Y_{i3} \\ R_{i1} & R_{i2} & R_{i3} \\ 0 & 0 & 1 \end{pmatrix}$.

$$
\begin{aligned}
& \mathcal{L}^{p7}_{i,obs} = \int\int \mathcal{L}_i dy_{i1} dy_{i2} \\
= \;& \int\int f(R_{i1}|y_{i1}) f(R_{i2}|R_{i1}, y_{i2}) f(y_{i1} f(y_{i2}|y_{i1}) f(y_{i3}|y_{i2}) dy_{i1} dy_{i2} f(R_{i3}|R_{i2}, y_{i3}) \\
= \;& \frac{\exp(\beta_1^{'} \mathbf{X_{i3}} + \boldsymbol{\theta_1} \mathbf{Y_{i3}})}{\sum_{s=1}^{2} \exp(\beta_s^{'} \mathbf{X_{i3}} + \boldsymbol{\theta_s} \mathbf{Y_{i3}})} (1-\pi_{i1}) \int\int \frac{\exp(\beta_2^{'} \mathbf{X_{i2}} + \boldsymbol{\theta_2} \mathbf{Y_{i2}})}{\sum_{s=1}^{2} \exp(\beta_s^{'} \mathbf{X_{i2}} + \boldsymbol{\theta_s} \mathbf{Y_{i2}})} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(\frac{-1}{2\sigma_1^2}(y_{i1} - \mu_{i1})^2) \\
\times \;& \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho_1^2)}} \exp(\frac{-1}{2\sigma_2^2(1-\rho_1^2)} (y_{i2} - \mu_{i2} - \rho_1 \frac{\sigma_2}{\sigma_1}(y_{i1} - \mu_{i1}))^2) \\
\times \;& \frac{1}{\sqrt{2\pi\sigma_3^2(1-\rho_2^2)}} \exp(\frac{-1}{2\sigma_3^2(1-\rho_2^2)} (y_{i3} - \mu_{i3} - \rho_2 \frac{\sigma_3}{\sigma_2}(y_{i2} - \mu_{i2}))^2) dy_{i1} dy_{i2} \\
= \;& \frac{\exp(\beta_1^{'} \mathbf{X_{i3}} + \boldsymbol{\theta_1} \mathbf{Y_{i3}})}{\sum_{s=1}^{2} \exp(\beta_s^{'} \mathbf{X_{i3}} + \boldsymbol{\theta_s} \mathbf{Y_{i3}})} (1-\pi_{i1}) \\
\times \;& \sum_{k=1}^{m} \sum_{l=1}^{m} \frac{w_k w_l}{\pi} \frac{\exp(\beta_2^{'} \mathbf{X_{i2}} + \boldsymbol{\theta_2}(\mu_{\mathbf{i2}} + \sqrt{\mathbf{2}}\rho_{\mathbf{1}}\sigma_{\mathbf{2}}\tau_{\mathbf{k}} + \sqrt{\mathbf{2}\sigma_{\mathbf{2}}^{\mathbf{2}}(\mathbf{1}-\rho_{\mathbf{1}}^{\mathbf{2}})}\tau_{\mathbf{l}}))}{\sum_{s=1}^{2} \exp(\beta_s^{'} \mathbf{X_{i2}} + \boldsymbol{\theta_s}(\mu_{\mathbf{i2}} + \sqrt{\mathbf{2}}\rho_{\mathbf{1}}\sigma_{\mathbf{2}}\tau_{\mathbf{k}} + \sqrt{\mathbf{2}\sigma_{\mathbf{2}}^{\mathbf{2}}(\mathbf{1}-\rho_{\mathbf{1}}^{\mathbf{2}})}\tau_{\mathbf{l}}))} \\
\times \;& \frac{1}{\sqrt{2\pi\sigma_3^2(1-\rho_2^2)}} \exp(\frac{-1}{2\sigma_3^2(1-\rho_2^2)} (y_{i3} - \mu_{i3} - \rho_2 \frac{\sigma_3}{\sigma_2}(\mu_{i2} + \sqrt{2}\rho_1\sigma_2\tau_k + \sqrt{2\sigma_2^2(1-\rho_1^2)}\tau_l - \mu_{i2}))^2)
\end{aligned}
$$

BIBLIOGRAPHY

AAPOR. Standard definitions: final dispositions of case codes and outcome rates for surveys (4th ed.). AAPOR, Lenexa, KS, 2006.

M. Alfo and A. Maruotti. A selection model for longitudinal binary responses subject to non-ignorable attrition. *Statistics in Medicine*, 28:2435–2450, 2009.

P. D. Allison. *Missing Data. Series:Quantitative Applications in the Social Sciences*. CA: Sage Publications, Inc, 2001.

R. M. Altman. Mixed hidden markov models: An extension of the hidden markov model to the longitudinal data setting. *Journal of American Statistical Association*, 102(477): 201–210, 2007. DOI: 10.1198/016214506000001086.

L. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Annals of Mathematical Statistics*, 41(1):164–171, 1970.

G. Cairncross, B. Berkey, E. Shaw, R. Jenkins, B. Scheithauer, D. Brachman, J. Buckner, K. Fink, L. Souhami, N. Laperierre, M. Mehta, and W. Curran. Phase iii trial of chemotherapy plus radiotherapy compared with radiotherapy alone for pure and mixed anaplastic oligodendroglioma: Intergroup radiation therapy oncology group trial 9402. *Journal Of Clinical Oncology*, 24(18):2707–2714, 2006. DOI: 10.1200/JCO.2005.04.3414.

B. J. Chen and X. H. Zhou. Non-homogeneous markov process models with informative observations with an application to alzheimer's disease. *Biometrical Journal*, 53(3):444–463, 2011. DOI: 10.1002/bimj.201000122.

B. J. Chen, G. Y. Yi, and R. J. Cook. Analysis of interval-censored disease progression data via multi-state models under a nonignorable inspection process. *Statistics in Medicine*, 29(11):1175–1189, 2010. DOI: 10.1002/sim.3804.

H. C. Chen, A. K. Manatunga, R. H. Lyles, L. M. Peng, and M. Marcus. Flexible modeling of longitudinal highly skewed outcomes. *Statistics in Medicine*, 28(30):3811–3828, 2009. DOI :10.1002/sim.3754.

D. Commenges, P. Joly, L. Letenneur, and J. F. Dartigues. Incidence and mortality of alzheimer's disease or dementia using an illness-death model. *Statistics in Medicine*, 23 (2):199–210, 2004.

B. Cooper and M. Lipsitch. The analysis of hospital infection data using hidden markov models. *Biostatistics*, 5:223–237, 2004.

D. Cox and N. Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91:729–737, 2004.

P. Diggle and M. G. Kenward. Informative drop-out in longitudinal data analysis. *Applied Statistics*, 43:49–93, 1994.

D. A. Dillman. *Mail and Internet Surveys: The Tailored Design Method, 2nd Edition.* New York: John Wiley and Sons, 2010.

S. Gao. A shared random effect parameter approach for longitudinal dementia data with non-ignorable missing data. *Statistics in Medicine*, 13:211–219, 2004. DOI: 10.1002/sim.1710.

G. Gong and F. Samaniego. Pseudo maximum likelihood estimation: theory and applications. *The Annals of Statistics*, 9:861–869, 1981.

C. Guihenneuc-Jouyaux, S. Richardson, and J. I. M. Longini. Modeling markers of disease progression by a hidden markov process: application to characterizing cd4 cell decline. *Biomatrics*, 56:733–741, 2000.

A. Guolo. Pseudo-likelihood inference for regression models with misclassified and mismeasured variables. *Statistica Sinica*, 21(4):1639–1663, 2011. DOI: 10.5705/ss.2010.065.

J. G. Ibrahim and G. Molenberghs. Missing data methods in longitudinal studies: a review. *Test*, 18(1):1–43, 2009. DOI: 10.1007/s11749-009-0138-x.

S. Ie Cessie, E. G. de Vries, C. Buijs, and W. J. Post. Analyzing longitudinal data with patients in different disease states during follow-up and death as final state. *Statistics in Medicine*, 28(30):3829–3843, 2009. DOI: 10.1002/sim.3755.

C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto. Multi-state markov models for disease progression with classification error. *The Statistician*, 52(2): 193–209, 2003.

W. Jank and J. Booth. Efficiency of monte carlo em and simulated maximum likelihood in two-stage hierarchical models. *Journal of Computational and Graphical Statistics*, 12(1): 214–229, 2003.

M. G. Kendall. *Rank correlation methods, 4th Edition.* Hodder Arnold, 1976.

K. Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data, 2nd Edition.* New York:Wiley, 1987.

Q. Liu and D. A. Pierce. A note on gauss hermite quadrature. *Biometrika*, 81:624–629, 1994.

R. J. MacKay. Estimating the order of a hidden markov mode. *Canadian Journal of Statistics*, 30:573–589, 2002.

F. P. P. M. Mackworth, N. Quality of life self-reports from 200 brain tumor patients: comparisons with karnofsky performance scores. *Journal of Neuro Oncology*, 14:243–253, 1992.

A. Maruotti. Mixed hidden markov models for longitudinal data: An overview. *International Statistical Review*, 79(3):427–454, 2011. DOI: 10.1111/j.1751-5823.2011.00160.x.

C. E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92:162–170, 1997.

A. Munoz, V. Carey, J. Schouten, M. Segal, and B. Rosner. A parametric family of correlation structures for the analysis of longitudinal data. *Biometrics*, 48:733–742, 1992.

R. H. Nagler, S. W. Gray, A. Romantan, B. J. Kelly, A. DeMichele, K. Armstrong, J. S. Schwartz, and R. C. Hornik. Differences in information seeking among breast, prostate, and colorectal cancer patients: Results from a population-based survey. *Patient Education and Counseling*, 81:S54–S62, 2010.

J. Nocedal and S. J. Wright. *Numerical Optimization, 2nd Edition*. Berlin, New York, 2006.

W. Parke. Pseudo maximum likelihood estimation: the asymptotic distribution. *The Annals of Statistics*, 14:355–357, 1986.

M. Parzen, S. R. Lipsitz, G. M. Fitzmaurice, J. G. Ibrahim, A. B. Troxel, and G. Molenberghs. Pseudo-likelihood methods for the analysis of longitudinal binary data subject to nonignorable non-monotone missing. *Journal of data science*, 5:1–21, 2007.

P. M. Philipson, W. K. Ho, and R. Henderson. Comparative review of methods for handling drop-out in longitudinal studies. *Statistics in Medicine*, 27(30):6276–6298, 2008. DOI: 10.1002/sim.3450.

M. Putt, J. A. Long, C. Montagnet, J. H. Silber, V. W. Chang, K. Liao, J. S. Schwartz, C. E. Pollack, Y. N. Wong, and K. Armstrong. Racial differences in the impact of comorbidities on survival among elderly men with prostate cancer. *Med Care Res Rev*, 66:409–435, 2009.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc IEEE*, 77:257–286, 1989.

H. Richard and M. Lynn. Serial correlation in unequally spaced longitudinal data. *Biometrika*, 77(4):721–732, 1990.

J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90:106–121, 1995.

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. J. Wiley and Sons, New York, 1987.

C. A. Sabin, D. Ashby, S. Richardson, C. GuihenneucJouyaux, F. HansfordMiller, J. F. Lawless, P. Yan, P. V. Bertrand, O. O. Aalen, C. Berzuini, C. Gobbi, A. Gigli, W. R. Gilks, D. DeAngelis, M. A. Newton, F. vandePol, M. R. Segal, P. J. Solomon, and J. M. G. Taylor. Markov chains with measurement error: estimating the true course of a marker of the progression of human immunodeficiency virus disease (with discussion). *Applied Statistics*, 45(3):275–309, 1996.

S. L. Scott, G. M. James, and C. A. Sugar. Hidden markov models for longitudinal comparisons. *Journal of American Statistical Association*, 100(470):359–369, 2005. DOI: 10.1198/016214504000001592.

S. K. Sinha, A. B. Troxel, S. R. Lipsitz, D. Sinha, G. M. Fitzmaurice, G. Molenberghs, and J. G. Ibrahim. A bivariate pseudolikelihood method for incomplete longitudinal binary data with nonignorable non-monotone missingness. *Biometrics*, 25:2784–2789, 2010.

A. Spagnoli, R. Henderson, R. J. Boys, and J. J. Houwing-Duistermaat. A hidden markov model for informative dropout in longitudinal response data with crisis states. *Statistics and Probability Letters*, 81(7):730–738, 2011. DOI: 10.1016/j.spl.2011.02.005.

R. Sutradhar, L. Barbera, H. Seow, D. Howell, A. Husain, and D. Dudgeon. Multistate analysis of interval-censored longitudinal data: Application to a cohort study on performance status among patients diagnosed with cancer. *Journal of Epidemiology*, 173(4): 468–475, 2010.

M. J. Sweeting, V. T. Farewella, and D. De Angelis. Multi-state markov models for disease progression in the presence of informative examination times: an application to hepatitis c. *Statistics in Medicine*, 29:1161–1174, 2009.

A. S. Tan, A. Bourgoin, S. W. Gray, K. Armstrong, and R. C. Hornik. How does patient-clinician information engagement influence self-reported cancer-related problems? *Cancer*, 117(11):2569–2576, 2011. DOI:10.1002/cncr.25804.

A. B. Troxel, D. P. Harrington, and S. R. Lipsitz. Analysis of longitudinal data with non-ignorable non-monotone missing values. *Appl. Statistics*, 47:425–438, 1998a.

A. B. Troxel, L. S. R., and D. P. Harrington. Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. *Biametrika*, 85(3):661–672, 1998b.

A. B. Troxel, S. R. Lipsitz, G. M. Fitzmaurice, J. G. Ibrahim, D. Sinhad, and G. Molenberghse. A weighted combination of pseudo-likelihood estimators for longitudinal binary data subject to non-ignorable non-monotone missingness. *Statistics in Medicine*, 29: 1511–1521, 2010.

R. Tsonaka, G. Verbeke, and E. Lesaffre. A semi-parametric shared parameter model to handle nonmonotone nonignorable missingness. *Biometrics*, 65:425–438, 2009.

R. Turner. Direct maximization of the likelihood of a hidden markov model. *Computational Statistics and Data Analysis*, 52(9):4147–4160, 2008. DOI: 10.1016/j.csda.2008.01.029.

Z. Uhry, G. Hedelin, M. Colonna, B. Asselain, P. Arveux, A. Rogel, C. Exbrayat, C. Guldenfels, I. Courtial, P. Soler-Michel, F. Molinie, D. Eilstein, and S. W. Duffy. Multi-state markov models in cancer screening evaluation: a brief review and case study. *Statistics Methods in Medical Research*, 19(5):463–468, 2010. DOI: 10.1177/0962280209359848.

C. Varin, R. N., and D. Firth. An overview of composite likelihood methods. *Statistica sinica*, 21(1):5–42, 2011.

A. P. Verbyla and B. R. Cullis. Modelling in repeated measures experiments. *Appl. Statist.*, 39:341–356, 1990.

M. M. Wall and R. Li. Multiple indicator hidden markov model with an application to medical utilization data. *Statistics in Medicine*, 28:293–310, 2009.

M. H. Wang, G. Cairncross, E. Shaw, R. Jenkins, B. Schethauer, D. Brachman, J. Buckner, K. Fink, L. Souhami, N. Laperriere, M. Mehta, and W. Curran. Cognition and quality of life after chemotherapy plus radiotherapy (rt) vs. rt for pure and mixed anaplastic oligodendrogliomas: radiation therapy oncology group trial 9402. *International Journal Of Radiation Oncology Biology Physics*, 77(3):662–669, 2010. DOI: 10.1016/j.ijrobp.2009.06.004.

L. R. Welch. Hidden markov models and the baum-welch algorithm. *IEEE inf. Theory Soc. Newsl.*, 53(4):1–13, 2003.