JOINT MODELING WITH CENSORED DATA AND GROUP-BASED TRAJECTORY CLUSTERING

by

Ching-Wen Lee

B.S. Public Health, Kaohsiung Medical University, Taiwan, 2002

M.S. Health Administration, National Yang Ming University, Taiwan, 2004

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

THE GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Ching-Wen Lee

It was defended on

July 9, 2013

and approved by

Dissertation Advisor: Lisa A. Weissfeld, Ph.D., Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Chien-Cheng Tseng, Sc.D., Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Chung-Chou H. Chang, Ph.D., Associate Professor
Departments of Medicine, Biostatistics, and Clinical and Translational Science
School of Medicine and Graduate School of Public Health
University of Pittsburgh

Mary Ganguli, M.D. M.P.H., Professor
Departments of Psychiatry, Neurology, and Epidemiology
School of Medicine and Graduate School of Public Health
University of Pittsburgh

JOINT MODELING WITH CENSORED DATA AND GROUP-BASED TRAJECTORY CLUSTERING

Ching-Wen Lee, PhD

University of Pittsburgh, 2013

ABSTRACT

Trajectories of data are collected in a variety of settings and offer insight into the evolution of a disease in the fields of biomedical, human genetic, and public health research. However, trajectories based on serum biomarkers are often subjected to censoring due to the low sensitivity of the bioassay used to measure the marker. A joint modeling approach incorporating a binary outcome and bivariate normal longitudinal markers which subject to left-censoring is proposed as a method to understand the relationship between two longitudinal outcomes and a binary outcome. The binary outcome is fitted by a logistic regression model, and the bivariate correlated longitudinal data are modeled using a linear mixed model. The binary outcome and bivariate measurements are then joined through the random coefficients that are present in both models. A clinical example from the GenIMS study is given. The public health significance is that the proposed method examined the relationship of the two censored longitudinal biomarkers and the binary outcome in a joint modeling approach which provided for direct inference on the effects of the two censored longitudinal marker measurements on the evolution of the disease outcome in public health.

Secondly, latent group-based trajectory modeling has been widely used to categorize individuals into several homogeneous trajectory groups. If there exist a small number of individuals who have unique trajectory patterns that are not similar to those observed in the rest of the population, the latent group-based trajectory modeling may end up identifying a larger

number of latent trajectory groups with several groups containing very few individuals. Further analysis treating latent groups as a covariate may then cause unstable estimates of standard errors. The second part of this dissertation applies the idea of the tight clustering method in the human genetic field into group-based trajectory analysis to classify latent trajectory groups that are more efficient, and to classify miscellaneous individuals or outliers whose trajectory patterns are dissimilar to the patterns in the rest of the population. We used the Bayesian information criterion as the criterion for model selection. A clinical example from the Normal Aging PiB study is provided. The public health relevance is that this innovative method is able to identify latent trajectory groups and outliers making it widely applicable in any public health setting where longitudinal trajectories are of interest.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Biomarkers offer an insight into the characteristics of a condition and the evolution of a disease in the fields of biomedical, medical, and human genetic research. Most importantly, biomarkers used for screening provide a relatively inexpensive opportunity to achieve an early diagnosis and prognosis, playing an important role in contributing to public health. For example, the prostate-specific antigen (PSA) level is used to aid in the detection of prostate cancer and testing of KRAS gene mutations is also used to aid in the detection of colon cancer by the National Comprehensive Cancer Network (NCCN). More and more studies have collected longitudinal biomarker measurements in order to better understand the mechanisms underlying disease progression. However, many assays are not sensitive enough to measure values either below a lower detection limit or above an upper detection limit, resulting in marker measurements being censored at either the lower or upper detection limits. This dissertation is related to longitudinal marker analysis and will be composed of two parts. The first part is to apply a joint modeling technique to see how disease mortality is affected by multiple longitudinal marker measurements which are subject to detection limits due to the sensitivity of the assay. The second part is to explore homogeneous latent groups of longitudinal marker trajectories through the application of the tight clustering technique developed by Tseng and Wong in 2005.

1.1 JOINT MODELING WITH CENSORED DATA

Longitudinal analysis has been expanded to characterize the relationship between marker profiles and a disease outcome through joint modeling techniques, especially when modeling longitudinal biomarker measurements combined with a binary or time-to-event outcome. The objectives are to explore the within-subject variabilities of the marker trajectories and to relate the features of the marker with a disease outcome [Tsiatis and Davidian, 2004]. The basic concept is to model the longitudinal measurements through a linear mixed effects model combined with the binary outcome through a generalized linear model or the event time through a survival model, that is to connect the longitudinal model and outcome model by shared random effects.

The censoring issue due to the sensitivity of an assay cannot be avoided. There are several ways to adjust for censored measurements. The common method is to use the imputation of a proportion of the detection limits, i.e. LOD or LOD/2 for the left-censored data. However, the naïve imputation approach has been shown to bias estimates and standard errors. Censored data can also be viewed as missing data and integrated out of the likelihood function [Hughes, 1999; Jacqumin-Gadda et al., 2000]. One can model left-censored data using a cumulative distribution function and right-censored data with a survival function. Either treating censored data as missing data or incorporating it into the model is shown to effectively adjust for the censoring issue [Thiebaut and Jacqmin-Gadda, 2004].

The first aspect of the dissertation is motivated by the Genetic and Inflammatory Markers of Sepsis (GenIMS) study. This prospective multicenter observational study recruited 2,320 community-acquired pneumonia (CAP) patients who were present in hospital emergency rooms and further 1895 confirmed CAP admitted for hospitalization to investigate the likelihood of

developing sepsis and its subsequent outcome, i.e. 90-day mortality. Demographics, examination results, and related information were collected and blood samples were collected during the first week of admission for all hospitalized subjects in the cohort. Researchers are interested in multiple biomarkers, including the inflammatory markers IL-6 and IL-10, and a coagulation/hemostasis marker D-dimer, and how these markers are simultaneously correlated with the disease outcome. Specifically, we modeled the disease outcome in a logistic regression model and the two biomarkers of interest, IL-6 and IL-10, in a bivariate linear mixed model. We used the covariance structure to capture the correlation between the two marker trajectories. However, due to the sensitivity of the assay, the markers were censored at either lower and/or upper detection limits. To better understand the relationship between marker evolutions and the disease outcome, we adjusted for the censored measurements by modeling left-censored data in a cumulative distribution function and right-censored data in a survival function. Our proposed nonlinear mixed model which combined two censored longitudinal profiles and a disease outcome was first fitted in the NLMIXED procedure in SAS 9.2. Simulation studies were also performed at various levels of censoring and different sample sizes. The simulation also compared our proposed method with the naïve imputation method.

1.2 GROUP-BASED TRAJECTORY CLUSTERING

There are several types of clustering methods available for identifying meaningful groups within a set of data. These methods include the hierarchical clustering algorithm [Eisen et al. 1998], the k-means algorithm [Tavazoie et al., 1999], self-organizing maps algorithm [Tamayo et al., 1999], tight clustering algorithm [Tseng and Wong, 2005], and model-based clustering that can be applied for data that are measured at a single point in time. For data that are collected

longitudinally, clustering methods have been extended to incorporate repeated measure structures and to classify the heterogeneity of marker trajectories, i.e. hierarchical clustering, growth curve model, mixtures of experts, regression mixtures, random effect regression mixtures, and mixtures of linear mixed models. To our knowledge, only the tight clustering algorithm distinguishes scattered data from tight clusters, and has not been applied to the developmental trajectory data structure.

The second aspect of this dissertation is inspired by the Normal Aging Pittsburgh Compound B (PiB) study. One goal of this study is to identify groups of subjects with similar longitudinal neuropsychological task trajectories and to understand the relationship between membership in a given trajectory class and the amount of amyloid deposition in the brain. The study recruited cognitive unimpaired elderly volunteers from the community and followed the participants for at least five years. All participants were examined by neuropsychological tests each year from recruitment. The neuropsychological tests included the N-back task and the letter-number sequencing task (used for working memory), the color-word Stroop test and the Hayling test (used for inhibitory efficiency) and so forth. All participants also received PiB positron emission tomography (PET) scanning, and magnetic resonance imaging (MRI). We applied the idea from the tight clustering technique [Tseng and Wong, 2005] to group similar trajectories of these neuropsychological scores together to better understand the developmental trajectories of the scores and to reduce the dimension of measurements. The proposed method is a repeated resampling process that is used to identify more efficient and similar trajectories into a group through the group-based trajectory method and to differentiate miscellaneous individuals that fit into none of the clustered groups. The longitudinal scores were modeled by linear random intercept models. The number of clusters was finite and assumed to follow a multinomial logistic

regression model. Simulation studies were conducted and the analysis was implemented in the FlexMix package in R.

2.0 JOINT MODELING OF A BINARY OUTCOME AND BIVARIATE LONGITUDINAL MARKERS SUBJECT TO CENSORING DUE TO DETECTION LIMITS

2.1 INTRODUCTION

There is currently a focus in medical research on identifying biomarkers that are predictive of prognosis and/or recovery from a given illness. This is a complex problem that occurs in many biomedical areas as it includes the handling of multiple biomarkers measured longitudinally coupled with outcomes that may be binary. To better understand the mechanisms driving disease, biomarkers are often measured from different disease pathways necessitating the development of statistical methods that can lead to an understanding of the relationship between pathways. Examples include the study of the relationship between CD4+ (T-lymphocytes) cell counts and plasma HIV RNA viral load and the contribution of this relationship to opportunistic diseases and/or death (Mellors et al., 1997). An additional, and motivating, example for this work came from a cohort study of Genetic and Inflammatory Markers in Sepsis (GenIMS). One goal of this study was to examine the relationship between pro- and anti-inflammatory biomarkers and in-hospital mortality for subjects with community acquired pneumonia (Kellum et al., 2007; Kale et al., 2010). In this case, there is a hypothesis that pro-inflammatory and anti-inflammatory biomarkers have competing trajectories, that is, as one marker increases, the other decreases;

making the joint consideration of these biomarkers extremely important in understanding risk of in-hospital mortality. This analysis is further complicated by the fact that the biomarkers are subject to left censoring due to the lower limit of detection of the assay measuring the marker.

To explore the relationship between one or more longitudinal profiles combined with a primary endpoint; a joint modeling approach has proven useful in many applications (Lin et al., 2002; Thiebaut et al., 2005). Under this framework, there are two submodels: a longitudinal submodel for the marker trajectories; and a survival or generalized linear submodel for the primary endpoint. The interrelationships between the longitudinal profiles and primary endpoint are formulated through the shared random effects (Li et al., 2004, 2007) or latent classes (Lin et al., 2002; Proust-Lima et al., 2007). Wang et al. (2000) showed that the naive two-stage method using least square estimates of the random coefficients of a linear mixed model as covariates of a logistic model was biased, and hence they modified the estimating equations based on regression calibration methods to improve the estimation. Later Wang and Huang (2001) proposed functional methods by extending the sufficiency score and conditional score estimators developed by Stefanski and Carroll (1987). Li et al. (2004) relaxed the distributional assumptions on the random effects of the longitudinal processes. Li et al. (2007) further proposed estimation procedures that require neither distributional or dependence structural assumptions on the random effects nor an independence assumption on the measurement errors. Bayesian approaches have also been developed for joint models with a binary endpoint (Horrocks et al., 2009). These aforementioned approaches assume that the target population is homogeneous and follows a single pattern of longitudinal profiles. The latent class model, on the other hand, assumes that a population is composed of various subpopulations with differing longitudinal

evolutions and that the influence of the evolution of these longitudinal markers on a disease outcome is different in each latent class (Lin et al., 2000; Proust-Lima et al., 2007 and 2009).

The methods discussed above are all useful for the joint analysis of multiple longitudinal measures coupled with a binary outcome. However, none of these methods address the issue of censoring in the longitudinal outcome. For the modeling of biomarker data, left-censoring is a common occurrence and further complicates the analysis in a joint modeling setting. The inflammatory markers, interleukin-6 (IL-6) and interleukin-10 (IL-10) measured in the GenIMS study were heavily left-censored due to the lower limits of detection of the assays used to measure these quantities. The censoring rate varied from 30% to 70% in these markers, necessitating the use of statistical methods that can accommodate this level of censoring. Ad hoc approaches to handling the censoring issue include substituting the value of the lower limit of detection or a fraction of the detection limit; however, this crude imputation leads to bias in the estimation of parameters and their standard errors (Hughes, 1999; Jacqmin-Gadda et al., 2000; Thiebaut and Jacqmin-Gadda, 2004). A common method to deal with censored data is to model left-censored data in a cumulative distribution function and right-censored data in a survival function (Thiebaut et al., 2005; Wannemuehler et al., 2010).

To simultaneously solve the issues of correlation and censoring, we propose a likelihood-based joint model for a binary outcome with an adjustment for the censored longitudinal covariate processes. Specifically, we construct a logistic regression model for the binary outcome and a bivariate linear mixed model for the two longitudinal markers by taking the censoring into account. We evaluate the proposed method and compare it to the naive

substitution methods through simulation studies. We illustrate the application of this method with an example from the GenIMS study.

2.2 METHOD

2.2.1 Modeling trajectories of bivariate longitudinal biomarkers through a mixed model

Repeated measurements of the two markers are modeled using a bivariate linear function of fixed and random effects:

$$y_{ik} = x_{ik}\beta_k + z_{ik}b_{ik} + \epsilon_{ik}. \quad (1)$$

Here, $y_{ik} = (y_{i1k}, \dots, y_{iT_ik})^T$, is a column vector of t_i time components for the k[th] biomarker readings in the i[th] subject. The $T_i \times p_k$ matrix of covariates for the k[th] biomarker is denoted as $x_{ik}^T = (x_{i1k}, \dots, x_{iT_ik})^T$, which may include polynomial terms of time. The t[th] row, x_{itk}^T is a p_k-dimensional vector of covariate values for biomarker k and for subject i. Note that $z_{ik}^T = (z_{i1k}, \dots, z_{it_ik})^T$ is a $T_i \times q_k$ matrix for the covariates of the random effects and that β_k and b_{ik} are p_k- and q_k- dimensional vectors of fixed and random coefficients, respectively. We assume that b_{ik} follows a multivariate normal distribution with mean zero and covariance structure, Δ. The off-diagonal components of Δ capture the correlation between the two markers and the serial correlation within a marker. The error term, ϵ_{ik}, is a T_{ik}-dimensional vector that is assumed to be uncorrelated with other variables, and to be normally distributed with mean zero and covariance

matrix $\sigma^2 I_{T_{ik}}$. We assume that the random effects and measurement error for each marker are independent of each other.

2.2.2 Modeling the binary outcome

Let R_i be an indicator of whether the i[th] subject experienced the disease outcome. It is natural to use simple logistic regression to model the probability of the primary outcome:

$$P(R_i = 1) = \frac{\exp(\eta)}{1 + \exp(\eta)}, \tag{2}$$

where η is the corresponding linear predictors $\eta = x_i\theta + b_i\gamma$.

We assume that the association between the two biomarker trajectories and the disease risk is represented by γ. The fully observed likelihood function can be denoted as

$$L(\theta, \beta, \gamma; r, y) = \prod_{i=1}^{N} \int f(r_i|\theta, \gamma, b_i) \prod_{k=1}^{2} f(y_{ik}|b_i, \beta_k) f(b_i) db_i \tag{3}$$

The first term of the right hand side of (3) is the likelihood function for the logistic model. The second term is for the bivariate linear mixed model. The random effects, b_i, in the mixed models are the shared parameters that link the two models. Given the random effects b_i, the repeated measurements are assumed to be independent within each marker and between the two markers. When there is censoring involved in the biomarker measurements, we introduced an indicator d_{ijk} to specify whether a measure of a biomarker is observed ($d_{ijk} = 0$), left-censored at L_k ($d_{ijk} = 1$), or right-censored at U_k ($d_{ijk} = 2$). Therefore, $\prod_{k=1}^{2} f(y_{ik}|b_i, \beta_k)$ in (3) becomes

$$\prod_{k=1}^{2} f(y_{ik}|b_i, \beta_k)$$

$$= \prod_{j=1}^{T_i} \prod_{k=1}^{2} f(y_{ik}|b_i, \beta_k)^{I(d_{ijk}=0)} F(L_k|b_i, \beta_k)^{I(d_{ijk}=1)} \{1 - F(U_k|b_i, \beta_k)\}^{I(d_{ijk}=2)}, \qquad (4)$$

$$\text{where } d_{ijk} = \begin{cases} 0, & \text{if observed} \\ 1, & \text{if left} - \text{censored} \\ 2, & \text{if right} - \text{censored} \end{cases}.$$

In equation (4), f(.|.), F(.|.), and 1-F(.|.) are the conditional probability density function, cumulative distribution function, and survival function, respectively. The integrated likelihood over the random effects can be maximized by a variety of optimization techniques. We utilized the adaptive Gaussian quadrature technique to approximate the integral of the random effects by a weighted sum over predefined abscissas for the random effects. With an appropriate number of quadrature points, an accurate approximation can usually be obtained (SAS Institute Inc., 2008). We utilized the trust-region optimization technique to carry out the maximization. The trust-region technique defines a hyperelliptic region with a certain radius around which a model function is trusted to be an accurate approximation of the likelihood; when an adequate model of the likelihood is found within the region then the region is expanded, and vice versa (SAS Institute Inc., 2008). When applying the method, we can obtain the initial values by separately fitting a linear mixed model for the longitudinal marker readings and a logistic regression model for the primary outcome. Parameters were estimated iteratively with the initial random effects estimated from the empirical Bayes estimates. Successful convergence results in maximum likelihood estimates and their approximate standard errors are based on the second derivative

matrix of the likelihood function. The maximum likelihood estimates have desirable asymptotic properties. i.e. consistency, asymptotic normality, and asymptotic efficiency. When the sample size is substantially large, bias becomes less important (Fitzmaurice et al., 2004; Brown and Prescott, 2006). The approximate standard errors of the nonrandom parameters were then used to compute corresponding t statistics, p-values, and confidence limits. Both of our fixed and random effects enter nonlinearly into the model. We fitted a nonlinear mixed model using SAS NLMIXED (SAS version 9.2) in the analysis.

2.3 SIMULATIONS

We carried out simulation studies to assess the performance of the proposed method and to compare it to the naïve imputation method. We also examined the sensitivity of our method to various sample sizes and different level of censoring proportions.

2.3.1 The proposed method vs. the naïve imputation method

For the simulation studies with 300 replications, we simulated two longitudinal marker measurements across time and an associated binary outcome. We simulated 7 measurements for each marker for each of the 600 subjects based on the bivariate linear mixed model described in equation (1) including a random intercept and slope for each marker individually. The time was set up to be from 1 through 7. For the first marker, the coefficient parameters for the intercept, b_{11}, and time, b_{12}, were 3 and 0.7, respectively. For the second marker, the coefficient parameters for the intercept, b_{21}, and time, b_{22}, were 4 and 0.6, respectively. The correlation

between these two markers was reflected by a covariance structure for the random intercepts and the random slopes. The variances (diagonal elements) of the covariance structure of the first marker's intercept and slope as well as the second marker's intercept and slope were set to be $\sigma^2_{b_{11}}$=1.5, $\sigma^2_{b_{12}}$=2, $\sigma^2_{b_{21}}$=1.3, and $\sigma^2_{b_{22}}$=1.8, respectively. The correlation coefficient, ρ, was common across the random effects and was set to be -0.2. In the logistic regression model, the fixed covariate, x, was generated from a uniform distribution(18, 60). The association between the marker trajectories and the binary outcome was specified through the subject-specific random intercepts and slopes γ in the logistic model described in equation (2). The coefficients of the intercept, x, and the four subject-specific random effects in the logistic model were set to be θ_0=-2, θ_1=-0.05, γ_{11}=-0.6, γ_{12}=0.8, γ_{21}=0.7, and γ_{22}=-0.9. With this setting, the average event rate was 24.1%. The censoring cut-off point for each marker was determined by the overall sample quantiles for each marker. We compare the results of our proposed method to that of the naive joint analysis. The naive joint analysis replaced the censored observations with the half value of the detection limits in the joint analysis.

The simulation results with 300 replications and 20% censoring in each marker are presented for two different approaches: our proposed joint analysis and a naïve joint analysis. The results show that the biases of our proposed method are smaller and that the coverage percentages are closer to the nominal value, 0.95, when compared to the naive imputation method. The estimated standard errors of the proposed method are closer to the empirical standard deviations (Table 2.1).

Table 2.1. The proposed method vs. the naïve imputation method

Parameter (true val.)	Proposed joint analysis					Naïve joint analysis[1]				
	Bias	Relative Bias Rate	SE mean	SD	CP%	Bias	Relative Bias Rate	SE mean	SD	CP%
Bivariate linear mixed model										
M1 int (3)	0.010	0.003	0.057	0.057	0.943	-0.293	-0.098	0.056	0.059	0.000
M1 slope (0.7)	-0.011	-0.015	0.060	0.060	0.937	0.220	0.314	0.046	0.045	0.000
M2 int (4)	0.005	0.001	0.064	0.065	0.957	-0.323	-0.081	0.063	0.068	0.010
M2 slope (0.6)	-0.004	-0.007	0.057	0.055	0.960	0.200	0.333	0.044	0.043	0.003
Logistic model										
Int (-2)	-0.040	0.020	0.231	0.231	0.967	0.215	-0.108	0.176	0.176	0.700
X (-0.05)	-0.002	0.040	0.012	0.012	0.930	0.006	-0.115	0.010	0.011	0.907
M1 random int (-0.6)	-0.011	0.018	0.157	0.156	0.970	0.163	-0.271	0.122	0.123	0.700
M1 random slope (0.8)	0.014	0.018	0.146	0.148	0.947	0.091	0.113	0.125	0.130	0.907
M2 random int (0.7)	0.039	0.056	0.230	0.226	0.967	-0.457	-0.652	0.133	0.138	0.110
M2 random slope (-0.9)	-0.008	0.009	0.150	0.136	0.970	-0.334	0.371	0.173	0.161	0.507
Covariance structure										
δ_{11}(1.5)	-0.044	-0.030	0.103	0.099	0.923	0.131	0.087	0.108	0.100	0.807
δ_{21} (-0.346)	0.009	-0.026	0.088	0.085	0.960	0.364	-1.052	0.063	0.060	0.000
δ_{22} (2)	0.019	0.009	0.137	0.143	0.953	-0.755	-0.377	0.073	0.074	0.000
δ_{31} (-0.279)	0.000	-0.001	0.088	0.089	0.960	-0.263	0.942	0.090	0.088	0.143
δ_{32} (-0.322)	0.009	-0.028	0.093	0.090	0.960	-0.005	0.017	0.072	0.071	0.947
δ_{33} (1.3)	0.002	0.002	0.138	0.136	0.953	0.328	0.252	0.141	0.132	0.343
δ_{41} (-0.329)	0.007	-0.021	0.080	0.077	0.960	0.014	-0.042	0.062	0.058	0.953
δ_{42} (-0.379)	-0.019	0.050	0.085	0.086	0.927	0.196	-0.517	0.050	0.048	0.027
δ_{43} (-0.306)	-0.012	0.039	0.097	0.099	0.963	0.357	-1.168	0.069	0.066	0.000
δ_{44} (1.8)	0.006	0.003	0.124	0.131	0.933	-0.673	-0.374	0.067	0.071	0.000

M1, Marker 1; M2, Marker 2; rand int, random intercept; δ_{ij} represents the ith row and jth column of the covariance structure (the order of the rows or columns: M1 int, M1 slope, M2 int, M2 slope); Relative Bias Rate = bias / true value; SE Mean, estimated standard error; SD, empirical standard deviation; CP%, coverage percentage.

[1]The naïve imputation method replaced the censoring observations with the half value of the detection limits.

2.3.2 Sensitivity of variety levels of marker censoring

Three parallel simulations with 20% censoring in both of the markers, 40% censoring in both of the markers, and 30% censoring in one marker plus 60% in the other were also conducted. The parameter settings in this section are different from those in the previous section while the model

remains the same. In the simulation studies of 500 replications, we simulated two longitudinal marker measurements across time and an associated binary outcome. For this simulation, we simulated 7 measurements for each marker for each of the 1500 subjects based on a bivariate linear mixed model as described in equation (1). We included a random intercept and slope for each marker individually. The time was set to range from 1 through 7. For the first marker, the coefficient parameters for the intercept, b_{11}, and time, b_{12}, were 3 and -0.5, respectively. For the second marker, the coefficient parameters for the intercept, b_{21}, and time, b_{22}, were 2 and -0.6, respectively. The correlation between these two markers was reflected by a covariance structure for the random intercepts and the random slopes. The variances (diagonal elements) of the covariance structure of the first marker's intercept and slope, as well as the second marker's intercept and slope, were set to be $\sigma^2_{b_{11}}$=4, $\sigma^2_{b_{12}}$=2.89, $\sigma^2_{b_{21}}$=0.25, and $\sigma^2_{b_{22}}$=0.36, respectively. The correlation coefficient, ρ, was common across the random effects, and was set to be -0.2. In the logistic regression model, the time independent covariate, x, was generated from a uniform(18, 60). The association between the marker trajectories and the binary outcome was specified through the subject-specific random intercepts and slopes γ in the logistic model described in equation (2). The coefficients of the intercept, x, and the four subject-specific random effects in the logistic model were set to be θ_0=-7, θ_1=1, γ_{11}=0.4, γ_{12}=2.7, γ_{21}=0.3, and γ_{22}=1.4. The censoring cut-off point for each marker was determined by the overall sample quantiles for each marker. With this setting, the average event rate was about 30%.

The results indicated that for smaller censoring percentage, the bias in the estimates was also smaller. In addition the estimates were more efficient (smaller variance estimates and the estimates of variance) and the coverage rates were close to 95%. When the censoring percentage

is 30% for one marker and 60% for the other, the results indicate that the method generally performs well (Table 2.2).

2.3.3 Sensitivity of different sample sizes

Three parallel simulations with sample sizes of 500, 1000, and 1500 subjects and 500 replications were also run for the proposed method. The censoring percentage was fixed at 20%. The parameter settings in this section are the same as those in Section 2.3.2.

The results from these simulations indicate that the bivariate longitudinal sub-model performs well generally among the range of sample sizes. However, the mortality sub-model performs better when the sample sizes are greater than 1000 (Table 2.3).

Table 2.2. Sensitivity of the marker censoring percentages

Parameter (true value)	20% censoring					40% censoring					30% censoring for M1 and 60% for M2				
	Bias	Relative Bias Rate	SD	SE mean	CP%	Bias	Relative Bias Rate	SD	SE mean	CP%	Bias	Relative Bias Rate	SD	SE mean	CP%
Bivariate linear mixed model															
M1 int (3)	0.000	0.000	0.057	0.056	0.944	0.011	0.004	0.058	0.060	0.958	-0.005	-0.002	0.059	0.058	0.950
M1 time (-0.5)	-0.001	0.002	0.045	0.045	0.950	-0.010	0.020	0.045	0.049	0.954	0.008	-0.015	0.044	0.047	0.958
M2 int (2)	0.000	0.000	0.022	0.022	0.952	0.000	0.000	0.024	0.024	0.954	0.004	0.002	0.028	0.028	0.950
M2 time (-0.6)	0.001	-0.002	0.015	0.016	0.966	0.001	-0.002	0.017	0.018	0.956	-0.002	0.003	0.020	0.022	0.952
Logistic model															
Int (-7)	-0.348	0.050	0.830	0.805	0.976	-0.458	0.065	0.948	0.935	0.976	-0.427	0.061	1.010	0.933	0.966
X (1)	0.049	0.049	0.116	0.111	0.974	0.063	0.063	0.134	0.130	0.976	0.064	0.064	0.141	0.130	0.968
M1 random int (0.4)	0.017	0.043	0.127	0.123	0.954	0.032	0.081	0.131	0.136	0.970	0.029	0.072	0.146	0.135	0.940
M1 random slope (2.7)	0.124	0.046	0.341	0.328	0.972	0.152	0.056	0.373	0.381	0.986	0.176	0.065	0.424	0.384	0.964
M2 random int (0.3)	-0.024	-0.079	0.738	0.710	0.974	-0.002	-0.006	0.896	0.828	0.976	0.009	0.029	0.929	0.922	0.998
M2 random slope (1.4)	0.033	0.023	0.393	0.383	0.954	0.083	0.059	0.437	0.429	0.966	0.098	0.070	0.474	0.455	0.958
Covariance structure															
δ_{11} (4)	0.007	0.002	0.163	0.168	0.960	-0.036	-0.009	0.172	0.176	0.948	0.029	0.007	0.168	0.177	0.956
δ_{22} (2.89)	-0.008	-0.003	0.114	0.113	0.960	0.023	0.008	0.138	0.137	0.958	-0.003	-0.001	0.131	0.124	0.938
δ_{33} (0.25)	0.000	0.002	0.025	0.025	0.946	0.000	0.000	0.030	0.028	0.946	-0.006	-0.025	0.033	0.033	0.940
δ_{44} (0.36)	0.000	0.000	0.013	0.014	0.958	0.001	0.004	0.016	0.016	0.968	-0.002	-0.005	0.020	0.020	0.942
tho (-0.2)	0.000	0.001	0.007	0.007	0.940	-0.001	0.007	0.007	0.008	0.952	-0.003	0.013	0.008	0.008	0.934
Measurement error															
errsq1 (0.64)	0.000	-0.001	0.012	0.012	0.952	0.000	0.001	0.014	0.014	0.964	0.000	0.000	0.012	0.013	0.962
errsq2 (0.49)	0.000	0.001	0.009	0.009	0.948	0.000	-0.001	0.011	0.011	0.956	0.001	0.002	0.013	0.013	0.950

M1, Marker 1; M2, Marker 2; rand int, random intercept; δ_{ij} represents the i^{th} row and j^{th} column of the covariance structure (the order of the rows or columns: M1 int, M1 slope, M2 int, M2 slope); Relative Bias Rate = bias / true value; SE Mean, estimated standard error; SD, empirical standard deviation; CP%, coverage percentage.

Table 2.3. Sensitivity of the sample sizes 500, 1000, and 1500

Parameter (true value)	500 subjects					1000 subjects					1500 subjects				
	Bias	Relative Bias Rate	SD	SE mean	CP%	Bias	Relative Bias Rate	SD	SE mean	CP%	Bias	Relative Bias Rate	SD	SE mean	CP%
Bivariate linear mixed model															
M1 int (3)	-0.006	-0.002	0.096	0.098	0.950	-0.003	-0.001	0.068	0.069	0.954	0.000	0.000	0.057	0.056	0.944
M1 time (-0.5)	0.001	-0.002	0.076	0.078	0.966	0.002	-0.004	0.057	0.055	0.942	-0.001	0.002	0.045	0.045	0.950
M2 int (2)	-0.001	0.000	0.037	0.037	0.954	0.000	0.000	0.026	0.026	0.950	0.000	0.000	0.022	0.022	0.952
M2 time (-0.6)	0.001	-0.002	0.027	0.028	0.958	0.002	-0.003	0.019	0.020	0.964	0.001	-0.002	0.015	0.016	0.966
Logistic model															
Int (-7)	-1.691	0.242	3.651	2.359	0.978	-0.597	0.085	1.178	1.089	0.982	-0.348	0.050	0.830	0.805	0.976
X (1)	0.245	0.245	0.513	0.331	0.972	0.086	0.086	0.166	0.151	0.968	0.049	0.049	0.116	0.111	0.974
M1 random int (0.4)	0.083	0.209	0.385	0.283	0.970	0.026	0.065	0.173	0.158	0.942	0.017	0.043	0.127	0.123	0.954
M1 random slope (2.7)	0.658	0.244	1.575	0.950	0.984	0.218	0.081	0.489	0.439	0.964	0.124	0.046	0.341	0.328	0.972
M2 random int (0.3)	0.045	0.148	2.036	1.690	1.000	-0.081	-0.269	1.031	0.938	0.986	-0.024	-0.079	0.738	0.710	0.974
M2 random slope (1.4)	0.344	0.245	1.522	0.922	0.970	0.078	0.056	0.521	0.491	0.956	0.033	0.023	0.393	0.383	0.954
Covariance structure															
δ_{11} (4)	0.011	0.003	0.284	0.291	0.970	0.004	0.001	0.197	0.206	0.972	0.007	0.002	0.163	0.168	0.960
δ_{22} (2.89)	-0.014	-0.005	0.205	0.195	0.932	-0.012	-0.004	0.137	0.138	0.958	-0.008	-0.003	0.114	0.113	0.960
δ_{33} (0.25)	0.000	-0.002	0.044	0.043	0.958	-0.001	-0.004	0.032	0.030	0.932	0.000	0.002	0.025	0.025	0.946
δ_{44} (0.36)	0.001	0.002	0.024	0.024	0.960	0.001	0.001	0.017	0.017	0.964	0.000	0.000	0.013	0.014	0.958
tho (-0.2)	0.000	0.002	0.013	0.013	0.954	0.000	0.002	0.009	0.009	0.952	0.000	0.001	0.007	0.007	0.940
Measurement error															
errsq1 (0.64)	0.000	0.000	0.020	0.021	0.958	0.000	0.000	0.014	0.015	0.962	0.000	-0.001	0.012	0.012	0.952
errsq2 (0.49)	0.000	0.000	0.016	0.016	0.956	0.000	0.000	0.011	0.011	0.944	0.000	0.001	0.009	0.009	0.948

M1, Marker 1; M2, Marker 2; rand int, random intercept; δ_{ij} represents the i^{th} row and j^{th} column of the covariance structure (the order of the rows or columns: M1 int, M1 slope, M2 int, M2 slope); Relative Bias Rate = bias / true value; SE Mean, estimated standard error; SD, empirical standard deviation; CP%, coverage percentage.

2.4 APPLICATION

We applied our proposed method to the GenIMS study to investigate the effects of the two correlated inflammatory markers IL-6 and IL-10 on 90-day mortality among community-acquired pneumonia patients who presented in hospital emergency rooms and were then admitted for hospitalization. The study recruited 2320 patients; among them, 1214 (52.33%) subjects are male, 1838 (79.22%) are white and 1586 (68.36%) are with Charlson Comorbidity Index (CCI) greater than zero. The 90-day mortality rate is 10.26%. After we excluded missing values in the two marker measurements in the analysis, the sample size was reduced to 1884. Among them, 52.07% of subjects are male, 80.79% are white and 72.51% are with Charlson Comorbidity Index (CCI) greater than zero. The mean (SD) of age is 67.3 (16.8). The 90-day mortality rate is 11.36%. Based on a previous analysis of the GenIMS data (Kellum et al., 2010), the markers IL-6 and IL-10 are associated with the 90-day mortality rate. However, the two markers both suffered from a high percentage of censoring. The censoring percentages of the first week for IL-6 and IL-10 range from 13.46% to 35.54% and from 46.87% to 78.91%, respectively. We will use our proposed method to estimate the effects of IL-6 and IL-10 on the 90-day mortality rate by accounting for the censoring in the two markers. Both of the markers were transformed to natural log scales in the analysis.

The results of our proposed model show that on average, IL-6 and IL-10 decrease over the first week. In the mortality model, older age, gender, and a Charlson Comorbidity Index (CCI) greater than zero are associated with higher risk of mortality. For the marker effects, the

association between IL-6 and 90-day mortality remains significant in the presence of IL-10, and vice versa. The magnitude of the coefficient for the intercept of IL-6 is 0.4, which demonstrates that a higher baseline measurement leads to a higher risk of mortality. The magnitude of the coefficient for the slope of IL-6 is 2.7, which indicates that marker readings of IL-6 with a slope above the mean trajectory (increasing IL-6 profiles or slow decreasing in IL-6) will lead to a higher risk of mortality. Similar patterns were found for IL-10 (Table 2.4).

Comparing the proposed method to the naïve imputation method, the results with p-values < 0.05 are the same for both methods. However, the proposed method shows that both the CCI effect in the longitudinal IL-6 model and the male effect in the IL-10 model are borderline significant. In the mortality model, the estimates and SEs of the random effects were smaller in the proposed model compared to the naïve imputation method which represents that the proposed method is more efficient (Table 2.4).

Table 2.4. Results from the GenIMS study

		Proposed method			Naïve imputation method		
Model	Parameter	Estimate	SE	P value	Estimate	SE	P value
Longitudinal IL-6	Int	3.229	0.164	<.0001	3.108	0.139	<.0001
	Slope	-0.451	0.014	<.0001	-0.358	0.011	<.0001
	Age	0.010	0.002	<.0001	0.009	0.002	<.0001
	Male	0.315	0.072	<.0001	0.238	0.061	0.000
	CCI > 0	-0.155	0.083	0.061	-0.094	0.070	0.184
Longitudinal IL-10	Int	1.992	0.171	<.0001	1.779	0.077	<.0001
	Slope	-0.598	0.029	<.0001	-0.147	0.006	<.0001
	Age	0.003	0.002	0.217	0.002	0.001	0.117
	Male	0.122	0.074	0.099	0.036	0.033	0.281
	CCI > 0	-0.038	0.084	0.653	-0.046	0.038	0.231
Mortality	Int	-7.795	0.553	<.0001	-7.835	0.553	<.0001
	Age	0.060	0.006	<.0001	0.062	0.006	<.0001
	Male	0.385	0.164	0.019	0.363	0.164	0.027
	CCI > 0	0.768	0.216	0.000	0.773	0.218	0.000
	Subject specific						
	IL-6 Int	0.417	0.071	<.0001	0.591	0.089	<.0001
	IL-6 Slope	2.665	0.540	<.0001	3.739	0.815	<.0001
	IL-10 Int	0.337	0.082	<.0001	0.669	0.167	<.0001
	IL-10 Slope	1.353	0.349	0.000	3.280	1.515	0.031
Covariance structure	δ_{21}	1.800	0.130	<.0001	1.001	0.067	<.0001
	δ_{31}	-0.655	0.041	<.0001	-0.541	0.030	<.0001
	δ_{32}	-0.450	0.033	<.0001	-0.192	0.015	<.0001
	δ_{41}	-0.376	0.046	<.0001	-0.191	0.014	<.0001
	δ_{42}	-0.565	0.055	<.0001	-0.188	0.010	<.0001
	δ_{43}	0.164	0.013	<.0001	0.051	0.003	<.0001
	δ_{11}	4.230	0.188	<.0001	3.725	0.151	<.0001
	δ_{22}	2.911	0.187	<.0001	1.244	0.053	<.0001
	δ_{33}	0.206	0.013	<.0001	0.130	0.007	<.0001
	δ_{44}	0.304	0.027	<.0001	0.041	0.003	<.0001

Note: CCI: Charlson Comorbidity Index.

Reference groups for gender and CCI are female and CCI=0, respectively.

2.5 DISCUSSION

The proposed method addressed the joint association between a disease outcome and two biomarker profiles accounting for the correlation between the two markers and subject to a restriction on the detection limits. The simulations were conducted to compare the proposed method with the naïve imputation method, and to assess the performance of the proposed method among a range of censoring levels, as well as among a range of sample sizes. The results show that our proposed method performs better than the naïve imputation of censoring in terms of parameter estimates and coverage percentages. The proposed method still performed well when the censoring levels increased to 40% for both markers with 1500 of sample size, although the smaller the censoring percentage, the smaller the biases, the more efficient, the closer the coverage rates to 95%. Even when the censoring percentage was 30% for one marker and 60% for the other with 1500 subjects, the results also performed generally well. The bivariate longitudinal sub-model performed well when the sample sizes ranged between 500 and 1500; while both of the biases and variances were reduced in the mortality sub-model when sample sizes were greater than 1000. The maximum likelihood estimates have desirable large sample or asymptotic properties, hence, bias becomes less important when the sample size is substantially larger than 1000 (Fitzmaurice et al., 2004; Brown and Prescott, 2006). Correspondingly, the results indicate that for higher levels of censoring, larger sample sizes are needed, that is, when the censoring levels are 20% for both markers, a sample size of 1000 is large enough. When the censoring levels of at least one marker are above 50%, a sample size of 1500 is needed. In summary, the proposed method provides an efficient way to see multiple censored marker

evolutions contribute to the progression of a disease. However, when the sample size is large, the computation may be time-consuming. The model is based on the assumption that the longitudinal marker profiles are homogeneous in the study; while in the future, we may test the homogeneity assumption of the marker evolutions and apply a joint model of a disease outcome combined with latent-class trajectories when appropriate.

3.0 THE USE OF TIGHT CLUSTERING TECHNIQUES FOR GROUP-BASED TRAJECTORY MODELING OF LONGITUDINAL DATA

3.1 INTRODUCTION

Developmental profiles of markers provide information on the progress of a disease. Nagin (2009) contends that charting and understanding developmental trajectories are among the most fundamental and empirically important research topics in the social and behavioral sciences and medical research today. Examples include tracing the temporal profiles of memory neuropsychological test scores for understanding pre-dementia memory declines and testing the prostate-specific antigen (PSA) level for aiding in early detection of the initial development of prostate cancer. Many of the longitudinal studies contain meaningful groups in the population that are heterogeneous profiles that do not depend on a single covariate, e.g. age or gender, and may act distinctively on the process of a disease (Nagin, 2009). Latent group based trajectory modeling has been developed to categorize homogeneous trajectory groups of the population. However, there are some issues in the application of the latent group based trajectory modeling when many groups are found and only small numbers of subjects are in some of the groups. Whether the small groups are homogeneous trajectory groups or randomly distributed cannot be discerned from the data. This issue cannot be addressed with regular latent group based trajectory modeling which forces all individuals into a group. Among clustering techniques

developed for genetic analyses, the tight clustering algorithm (Tseng and Wong, 2005) provides a way to select clusters that are more stable and consistent. In this paper, we apply the idea from the tight clustering algorithm to the latent group based trajectory modeling to search for trajectory groups that are less variant and more homogeneous.

A variety of non-model-based heuristic clustering methods have been widely used in the literature, including hierarchical clustering (Eisen et al. 1998), self-organizing maps (Tamayo et al., 1999), K-means (Tavazoie et al., 1999), and the tight clustering algorithm (Tseng and Wong, 2005). Model-based clustering methods assume that the data are generated by a finite mixture of underlying probability distributions in which each component represents a different group or cluster; these are known as finite mixture models (Fraley and Raftery, 1998). For example, in the Gaussian mixture model, each cluster is modeled by the multivariate normal distribution. Gaussian mixture models have been shown to produce higher quality clustering results than heuristic approaches when the data are appropriately transformed (Yeung, et al., 2001). However, less attention has been paid to the setting where the data are measured across time. To differentiate temporal profiles of dysfunction levels on human health provides insights into how a disease evolves. Clustering array data with repeated measurements has been explored in recent years based on several different approaches to the problem (Yeung, et. al. 2003). Finite mixture models with a fixed number of components have been extended for such purposes. Some researchers proposed trajectory clustering with mixtures of regression models and mixtures of non-parametric regression models (Gaffney and Smyth, 1999; Grün et al., 2012). A mixture of standard linear regression models also falls under the umbrella of latent class regression (Leisch, 2004). Some researchers have used mixture models in the framework of mixed-effects models which incorporate error estimates estimated from repeated measurements (Yeung et al, 2003;

Celeux et al., 2005; Qin and Self, 2006; Grün and Leisch, 2008; Ma et al, 2009). Luan and Li (2003) proposed a mixture of mixed-effects models using B-splines to smooth the cluster profiles. Some clustering methods developed variability-weighted similarity measures that would down-weight noisy genes or noisy experiments (Yeung et al., 2003). Bayesian mixture-model-based clustering has been developed, e.g. with the structure of generalized linear models with random effects (Lenk and DeSarbo, 2000).

However, in most clustering studies, researchers classify each of the units i.e., subjects, into a cluster (Yeung, et al. 2001; Yeung, et al. 2003; Luan and Li 2003; Qin and Self, 2006), with the case being that either a large number of clusters are formed, or fewer clusters which contain small groups of miscellaneous individuals. In the former case when clusters with a small number of individuals are identified, analyses using cluster membership as a covariate may be unstable. Furthermore, there are some situations where some units are miscellaneous, i.e., their characteristics are different from the cluster that the individual was assigned (Tseng and Wong, 2005; Yuan and Kendziorski, 2005). When miscellaneous units are forced into a cluster, the estimation of the number of clusters, the parameters of the clusters, and the inference based on the parameters will be biased. We utilize the concept of tight clustering method which is used in microarray analysis (Tseng and Wong, 2005), and further apply it to repeated measures data. Using this approach, we develop a method that can classify trajectories into several homogeneous groups and also identify trajectories that do not belong to any of the homogeneous groups (the miscellaneous group). This allows us to distinguish units that do not contain useful information and to focus on the more efficient trajectories, namely the “cores” of the group, for estimation and inference. The paper is organized as follows: a detailed statistical method is

presented in section 2, followed by simulation studies in section 3, and an empirical application in section 4; we further give a discussion in the last section.

3.2 METHOD

We propose to combine the tight clustering technique and latent group-based trajectory method in order to find the latent trajectory groups and to identify miscellaneous individuals or outliers, i.e., trajectory patterns that are deviant from the rest of the population and/or trajectory patterns that are dissimilar to the patterns in the rest of the population. The tight clustering technique has been used in the microarray analysis which is an unsupervised learning and resampling method. The latent group-based trajectory method has been used to classify individuals into several homogeneous trajectory groups. We assume that 1) individuals from the same trajectory group share common effects of covariates, and 2) some individuals do not belong to any homogeneous trajectory group, e.g., miscellaneous individuals or outliers.

The number of groups is determined through minimizing the Bayesian information criterion (BIC; Schwarz, 1978; Brown and Prescott, 2006). The BIC is defined as $\log(L) - 0.5k\log(N-p)$, where L is the likelihood; p is the number of fixed effects in the model; N is the number of observations; and k is the number of covariance parameters.

3.2.1 Modeling the probability of group membership

Suppose there are N subjects in the population, $i = 1, \dots, N$, that can be divided into K latent groups, $k = 1, \dots, K$ and that ψ_{ik} represents the individual probability that the i^{th} subject

belongs to the latent group *k*. For each subject *i*, $\sum_{k=1}^{K} \psi_{ik} = 1$. Then ψ_{ik} can be modeled using the multinomial logistic regression as the form

$$\psi_{ik} = p(c_{ik} = 1|X_{1i}) = \frac{\exp(\rho_{0k} + X_{1i}^T \rho_{1k})}{1 + \sum_{j=2}^{K} \exp(\rho_{0j} + X_{1i}^T \rho_{1j})}, \forall k = 1, \dots, K.$$

For group *k*, ρ_{0k} denotes the intercept and ρ_{1k} denotes the $p_1 \times 1$ fixed parameters for covariates X_{1i}^T. For identifiability, we assume that the coefficients of the first group will be set to zero ($\rho_{01} = 0$ and $\rho_{11} = 0$).

3.2.2 Modeling trajectories of a longitudinal marker through a normal mixture model

We assume that the repeated measures of a subject belong to the same group. Once the number of groups is fixed, the marker measurements, y_i, are related to the latent group through mixture components which could be random effects, fixed effects, and/or some other covariates. For simplicity, the repeated measures, y_i, are represented as mixtures of linear random intercept models with each mixture component coming from a latent group, *k*. That is, given a latent group *k*, the repeated measurements of a marker are modeled as a linear normal random intercept model

$$y_{i|c_{ik}=1} = X_{2i}^T \beta_k + b_{ik} + \epsilon_i,$$

where $y_i = y_{it} = (y_{i1}, \dots, y_{im_i})^T$ is a column vector of longitudinal measurements of the i^{th} subject, $i = 1, \dots, N$, on the t^{th} measurement, $t = 1, \dots, m_i$, and X_{2i}^T is a design matrix associated with the p_2 vector of class-specific fixed effects β_k. The random intercept effect, b_{ik}, and the measurement errors, ϵ_i, are assumed to be independent as well as normally and identically distributed with mean 0 and variance $\sigma_{b_k}^2$ and σ_ϵ^2, separately. For simplicity, the measurement error variances σ_ϵ^2 are common to all clusters.

3.2.3 The log-likelihood of the joint model

Let θ denotes all of the parameters that are to be estimated, that is $\theta = \{\rho_{0k}, \rho_{1k}, \beta_k, b_{ik}, \sigma_{b_k}^2, \sigma_\epsilon^2\}$. Once the number of latent groups is fixed, the log-likelihood of the latent joint model given the observed data, $l(\theta; y)$, can then be specified as

$$\sum_{i=1}^{N} \ln\left\{\sum_{k=1}^{K} \frac{\exp(\rho_{0k} + X_{1i}^T \rho_{1k})}{\sum_{j=1}^{K} \exp(\rho_{0j} + X_{1i}^T \rho_{1j})} \times f_k(y; \theta)\right\},$$

where $\frac{\exp(\rho_{0k} + X_{1i}^T \rho_{1k})}{\sum_{j=1}^{K} \exp(\rho_{0j} + X_{1i}^T \rho_{1j})}$ calculates the individual probability of belonging to the latent class *k*; that is ψ_{ik}. In each class, $f_k(y; \theta)$ captures the longitudinal evolution of the marker.

$$f_k(y; \theta) = \prod_{j=1}^{m_i} f_k(y_j; \theta) = \frac{1}{(2\pi)^{\frac{m_i}{2}} |\sigma_\epsilon^2 I|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\{y - E(y)\}^T (\sigma_\epsilon^2 I)^{-1} \{y - E(y)\}\right].$$

3.2.4 Using the tight clustering algorithm

In the following steps, we will define groups that are less variant by using the idea of the tight clustering algorithm through a resampling method (Figure 3.1). We first fix the number of groups, *k*. Suppose we have a sample *Y* which is composed of *N* sampling units with *T* repeated measurements. In the resampling method, take a random subsample of *Y*, Y', as 70% of *Y*. This sampling is repeated and two subjects are classified into a given group based on the number of times that they fall within the same group. We estimate the parameters using the maximum likelihood estimation based on the EM algorithm (Dempster et al, 1977) on the subset of the sample and then find the mean trajectories and form the classifier $C(Y', k)$ to cluster the original data *Y* according to their posterior probabilities. The posterior probabilities can be specified as

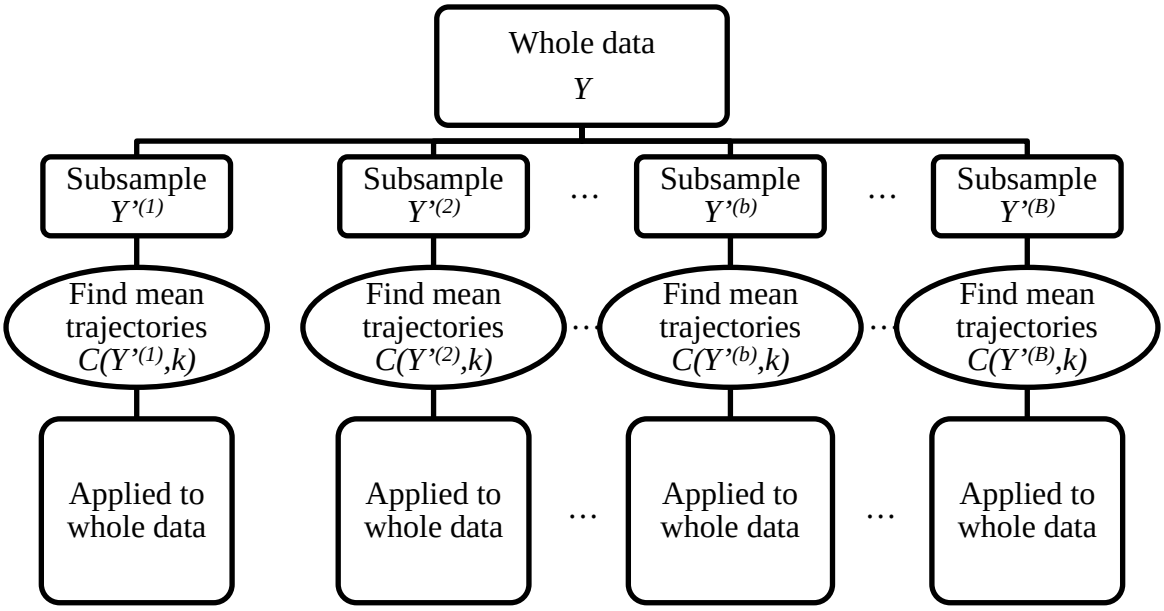

Figure 3.1. The re-sampling method

$p(c_{ik}=1|y_i;\hat{\theta}) = \frac{p(c_{ik}=1)f(y_i|c_{ik}=1;\hat{\theta})}{\sum_{j=1}^{K} p(c_{ij}=1)f(y_i|c_{ij}=1;\hat{\theta})}$. The result is represented by a $N \times N$ co-membership matrix $D[C(Y',k),Y]$, in which 1 represents subjects that are in the same group while 0 denotes different groups. The next step is to repeat the independent random sub-sampling *B* times to obtain subsamples $Y'^{(1)}, Y'^{(2)}, \ldots, Y'^{(B)}$. Following the same procedure, we calculate the average co-membership matrix by taking the mean of the all co-membership matrices. The search for a set of trajectories $V = \{v_1, \ldots, v_m\} \subset \{1, \ldots, n\}$ such that $\bar{D}_{v_i v_j} \geq 1 - \alpha, \forall i, j$ where α is a constant close to zero. These *V* sets are groups and ordered by size from largest to smallest. On the other hand, for the regular trajectory method as a comparison method, we estimate parameters based on the whole sample, *Y*. The maximum likelihood estimation is also based on the expectation-maximization (EM) algorithm. Then we cluster Y into a group according to their largest posterior probabilities of being classified in that group. We use the FlexMix package

which provides infrastructure for flexible fitting of finite mixture models in the R computing environment using the EM algorithm or its variants (Grün and Leisch, 2008). We report our results in two steps. Step 1 is to differentiate the miscellaneous individuals from the rest of the population. Step 2 is to classify the non-miscellaneous individuals into groups. Furthermore, we compared our proposed method with a regular latent group-based trajectory method using the FlexMix package accounting for random intercepts. In the regular method, once the number of groups is fixed, individuals are classified into a group with the largest posterior probability.

3.3 SIMULATION

To evaluate the performance of the proposed method, we conducted two simulation scenarios: two trajectory groups plus one miscellaneous group and three trajectory groups plus one miscellaneous group. We used two steps to assign individuals into a group. The first step was to determine a miscellaneous group and the next step was to classify the non-miscellaneous individuals by their underlying groups.

3.3.1 Two trajectory groups plus one miscellaneous group

In 100 simulations of the first scenario, we generated a linear random intercept model for each of the two trajectory groups (n = 50 for each group) and one miscellaneous group (n = 4). Each trajectory includes observations recorded at 7 time points. A covariate was generated from a uniform(18, 60). Specifically, the two trajectory groups with a sample size of 50 were

$$Y_{1im_i} = (0.5 + b_{1i}) + 1 \times time_{im_i} + 2 \times x_i + \epsilon_{im_i}, and$$

$$Y_{2im_i} = (1.5 + b_{2i}) - 0.5 \times time_{im_i} + 0.5 \times x_i + \epsilon_{im_i}.$$

The random intercepts were assumed to be normally and identically distributed with (mean, variance) of (0, 0.49) and (0, 0.81) for the two groups, respectively. The errors were assumed to be common across the two groups with mean 0, and variance 0.36. For the miscellaneous group of size 4, two observations were generated from a similar setting as that of the first trajectory group, but with the variances of the random intercept and errors of 12.1 and 9, separately, and another two were generated from a uniform(-80, 80).

To perform the resampling technique for each dataset, we randomly selected 70% of the sample without replacement 100 times. For simplicity, we used the Bayesian information criteria for the first dataset to provide an estimate of the number of groups. Once the number of groups was determined, we then estimated the parameters of each group. After the parameters were estimated using the FlexMix package, we applied the parameters back to the whole dataset so that a posterior probability of belonging to a specific group can be estimated for each individual of the dataset. An individual was assigned to a specific group if its corresponding posterior probability of that group was greater or equal to 0.8. An individual was assigned to the miscellaneous group if all of its posterior probabilities were less than 0.8 for that group. We then calculated the co-membership matrix for each whole sample, and derived the average co-membership by pulling all 100 subsamples in the resampling setting. To do this we searched for a set of trajectories such that any pair of average co-membership within the set was $\geq 1 - \alpha$, where α was set to 0.05. As a result, the set was selected as a trajectory group. A miscellaneous group was decided by that the size of the group was very small, i.e., less than 10 individuals in that group.

The results showed that our proposed method correctly assigned individuals into their underlying groups with high probability. In the first step, we calculated the mean ± SD of the rate that individuals from the miscellaneous group being assigned correctly to this group was 97.25% ± 7.86%. Its range was from 75% to 100% with a median rate of 100%. On the other hand, the overall mean ± SD was 99.50% ± 1.37% of the rate that individuals from the two trajectory groups were correctly not assigned to the miscellaneous group with a range of 94% to 100% and a median rate of 100%. In the second step only the subjects that were not generated as miscellaneous subjects were included. The mean ± SD was 100% ± 0% of the rate that individuals from the two trajectory groups being correctly assigned to their latent groups.

Table 3.1 was composed to compare the mean (SD) of the estimates from our proposed method with the regular method by groups. Although the mean estimates of the intercept and the variance of the random intercept were biased, the mean estimates of the rest variables in the proposed method were closer to the true values and more efficient compared to the regular method.

Table 3.1. Mean (SD) of the estimates

	Group 1			Group 2		
	True value	Proposed method	Regular method	True value	Proposed method	Regular method
Int	0.5	1.7 (2.39)	0.79 (1.60)	1.5	0.27 (2.57)	1.51 (1.71)
Time	1	1 (0.02)	0.73 (0.39)	-0.5	-0.5 (0.02)	-0.26 (0.39)
Age	2	1.85 (0.25)	1.74 (0.39)	0.5	0.64 (0.24)	0.76 (0.41)
σ^2_{error}	0.36	0.35 (0.03)	39.93 (39.94)	0.36	0.35 (0.03)	47.91 (37.78)
σ^2_{rint}	0.49	18.79 (33.52)	37.07 (55.34)	0.81	19 (33.80)	37.53 (54.68)

3.3.2 Three trajectory groups plus one miscellaneous group

In 100 simulations of the second scenario, we generated a linear random intercept model for each of the three trajectory groups (n = 40, 30, and 30 for each group respectively) and one miscellaneous group (n = 4). Each trajectory includes observations recorded at 7 time points. A covariate was generated from a uniform(18, 60). Specifically, the three trajectory groups with a sample size of 40, 30, and 30, separately, were

$$Y_{1im_i} = (0 + b_{1i}) + 0.5 \times time_{im_i} + 2 \times x_i + \epsilon_{1im_i},$$

$$Y_{2im_i} = (1.5 + b_{2i}) - 1 \times time_{im_i} + 0.5 \times x_i + \epsilon_{2im_i}, and$$

$$Y_{3im_i} = (2 + b_{3i}) + 4 \times time_{im_i} + 1 \times x_i + \epsilon_{3im_i}.$$

The random intercepts were assumed to be normally and identically distributed with (mean, variance) of (0, 0.81), (0, 0.49) and (0, 0.25) for the three groups, respectively. The errors were assumed to be common across the two groups with mean 0, and variance 0.36. As to the miscellaneous group of size 4, two of the subjects were generated from a similar setting as that of the first trajectory group but with the variances of the random intercept and errors of 12.1 and 9, separately, and another two were generated from a uniform(-80, 80).

To perform the resampling technique for each dataset, we randomly selected 70% of the sample without replacement 100 times. For simplicity, we used the Bayesian information criteria for the first dataset to provide an estimate of the number of groups. Once the number of groups was determined, we then estimated the parameters of each group. After the parameters were estimated using the FlexMix package, we applied the parameters back to the whole dataset so that a posterior probability of belonging to a specific group can be estimated for each individual

of the dataset. An individual was assigned to a specific group if its corresponding posterior probability of that group was greater or equal to 0.8. An individual was assigned to the miscellaneous group if all of its posterior probabilities were less than 0.8 for that group. We then calculated the co-membership matrix for each of the whole sample, and derived the average co-membership by pulling all 100 subsamples in the resampling setting. To do this we searched for a set of trajectories such that any pair of average co-membership within the set is $\geq 1-\alpha$, where α is set to 0.05. As a result, the set was selected as a trajectory group. A miscellaneous group was decided by that the size of the group was very small, i.e., less than 10 individuals in that group.

The results showed that our proposed method correctly assigned individuals into their underlying groups with a high probability. In the first step, we calculated the mean ± SD, 99.50% ± 3.52%, representing the rate that individuals from the miscellaneous group were correctly assigned to this group. Its range was from 75% to 100% with a median rate of 100%. On the other hand, the overall mean ± SD was 98.82% ± 3.31% of the rate that individuals from the three trajectory groups were correctly not assigned to the miscellaneous group with a range of 82% to 100% and a median rate of 100%. In the second step only the subjects that were not generated as miscellaneous subjects were included, the mean ± SD was 98.71% ± 6.38%, 98.52% ± 7.29%, and 99.85% ± 1.49%, respectively, of the rate that individuals from the three trajectory groups being correctly assigned to their latent groups.

Table 3.2 compared the mean (SD) of the estimates from our proposed method with the regular method by three groups. As in Table 3.1, although the mean estimates of the intercept and the variance of the random intercept were biased, the mean estimates of the rest variables in

the proposed method were closer to the true value and more efficient compared to the regular method.

Table 3.2. Mean (SD) of the estimates

	Group 1			Group 2			Group 3		
	True value	Proposed method	Regular method	True value	Proposed method	Regular method	True value	Proposed method	Regular method
Int	0	0.6 (2.83)	0.63 (1.93)	1.5	1.83 (1.67)	1.09 (3.30)	2	0.64 (2.37)	1.65 (1.98)
Time	0.5	0.47 (0.14)	0.76 (1.05)	-1	-0.3 (1.75)	0.08 (1.74)	4	3.38 (1.63)	2.9 (1.68)
Age	2	1.92 (0.20)	1.75 (0.42)	0.5	0.53 (0.18)	0.72 (0.38)	1	1.06 (0.14)	1.09 (0.26)
σ^2_{error}	0.36	0.45 (0.51)	18.18 (41.74)	0.36	0.35 (0.03)	85.71 (82.30)	0.36	0.42 (0.42)	44.82 (67.43)
σ^2_{rint}	0.81	11.94 (31.13)	24.75 (49.28)	0.49	5.23 (14.06)	28.08 (64.04)	0.25	13.17 (24.07)	29.04 (50.13)

3.4 APPLICATION

We applied our method to the Normal Aging Pittsburgh Compound B (PiB) study to identify latent groups of longitudinal neuropsychological measurements excluding a group of outliers. One goal of this study is to identify groups of subjects with similar longitudinal neuropsychological task trajectories and to understand the relationship between membership in a given trajectory class and the amount of amyloid deposition in the brain. The 79 elderly volunteers without cognitive impairments were recruited from the community through an advertisement in a local seniors' newspaper and through direct mailings to individuals who had shown an interest in participating. The exclusion criteria included the presence of dementia or mild cognitive impairment (MCI), the presence or history of major neurological or psychiatric

diseases, and the use of psychoactive medications at the time of recruitment (Aizenstein et al., 2008). The study followed participants for at least 5 years. Participants were given a variety of neuropsychological tests at recruitment and at annual follow-ups. The neuropsychological tests included the N-back task and a letter-number sequencing task (used for working memory), the color-word Stroop test and the Hayling test (used for inhibitory efficiency), and among other applicable tests. All participants also received PiB positron emission tomography (PET) scanning and magnetic resonance imaging (MRI). The 79 subjects had at least 12 years of education and had average (SD) age 75.8 (6.2) years old (range 65 - 92 years old). Among these subjects, 28 (35.4%) were male, 68 (86.1%) were white, and 14 (19.2%) were APOE*4 allele carriers. In the analysis, we used our proposed method to identify latent trajectory groups of N-Back task measurements (baseline and 4 follow-ups) and to identify individuals whose trajectory of N-Back task measurements outsides of these patterns. Based on the BIC, two latent trajectory groups were determined, group 1 and group 2 with a size of 35 and 23, respectively. In general, the first group started with a high N-Back task score and increased slightly over five years, while the second group started with a lower score and increased faster than the first group. There were 21 subjects with miscellaneous trajectories that were represented as group 3. We compared the results obtained from our method to those obtained from a regular latent group-based trajectory method using FlexMix accounting for random intercepts. The spaghetti plot (Figure 3.2) illustrated the N-Back task measurements over time for the two groups (group 1 and group 2), and the rest of the individuals (a miscellaneous group, group 3). We then made a cross-table of the results from our method with results from the regular latent group-based trajectory method (Table 3.3). Our method identified that, when using the regular latent trajectory analysis, 14

miscellaneous subjects were included in group 1 and 7 miscellaneous subjects were included in group 2.

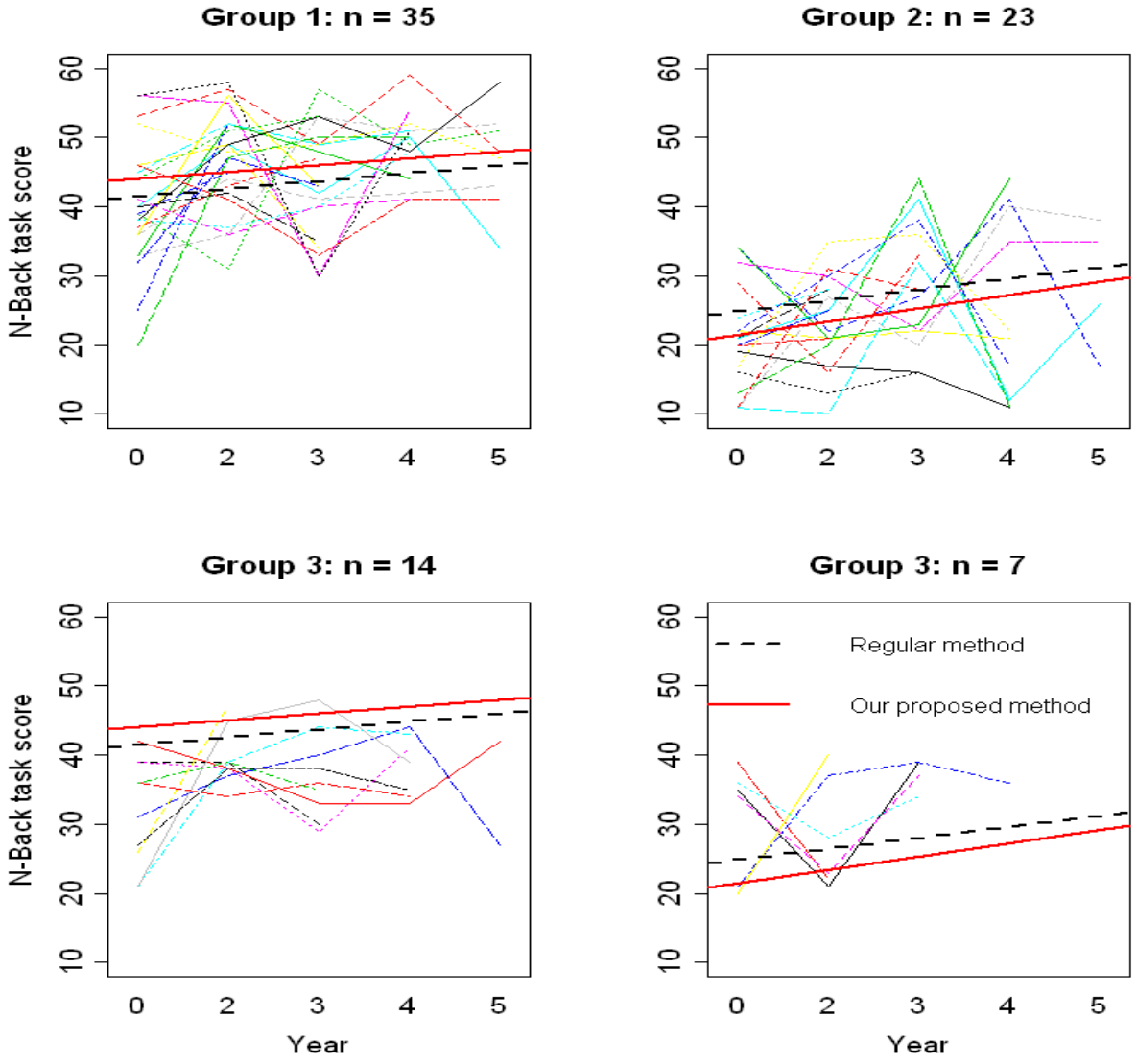

Figure 3.2. Spaghetti plots of individual profiles by the latent trajectory groups

Note. Fitted latent trajectories of the groups were obtained from our proposed method (red lines) and from the regular latent group-based trajectory method (black dashed lines).

Table 3.3. The crosstab of groups by our proposed method and the regular method

	Our proposed method			
Regular method	Group 1	Group 2	Miscellaneous	Total
Group 1	35	0	14	49
Group 2	0	23	7	30
Total	35	23	21	79

We then compared the demographic characteristics between the 35 subjects in group 1 by our proposed method and 14 miscellaneous subjects who were classified into group 1 by the regular method (Table 3.4). The results presented that age, race, and APOE*4 allele carriers showed a borderline significant difference between the two groups. The 35 subjects in group 1 classified by our proposed method were younger and contained more white people and less APOE*4 allele carriers than the 14 miscellaneous subjects. When comparing the demographic characteristics of between the 23 subjects in group 2 identified by our proposed method and 7 miscellaneous subjects who were classified into group 2 by the regular method (Table 3.4), the results showed that none of the demographics were significantly different between the two groups. Although there were differences in the percentages of the APOE*4 allele carriers between the two groups, it was not statistically significant; this was probably due to a power issue.

3.5 DISCUSSION

Our proposed method successfully identifies latent trajectory groups of longitudinal measurements that are more efficient (less variant) and to identify the miscellaneous group. Once the trajectory groups were found, researchers can use these group memberships for further

Table 3.4. The demographic characteristics comparisons

Demographics	Group 1 $n_1 = 35$	Group 3 $n_3 = 14$	P value	Group 2 $n_2 = 23$	Group 3 $n_3 = 7$	P value
Mean age (SD)	71.9 (5.5)	74.9 (5.9)	0.077	75.3 (6.4)	75.6 (4.4)	0.884
Mean education (SD)	15.1 (2.8)	15.2 (2.8)	0.937	14.6 (2.2)	14.3 (2.7)	0.638
Male, %	22.9	42.9	0.181	43.5	42.9	0.999
White, %	91.4	71.4	0.091	87.0	85.7	0.999
APOE*4, %	12.1	38.5	0.092	20.0	14.3	0.999

analysis. Clinical inspection may be required for individuals in the miscellaneous groups in order to provide subject-specific health care. The value of α used as a threshold for the pair of the level of the average comembership in a group ensures the levels of group stability. Our simulation results of the two trajectory-group and the three trajectory-group studies showed a high probability of correctly identifying individuals in the miscellaneous group. However, when the sample size is large, computations may be time-consuming. In the future, we will extend the study for a large sample size, investigate the effects of the magnitude of α on the results, and examine the influences of the proportion levels of the subsamples taken in the resamping method.

BIBLIOGRAPHY

Abbas OA. (2008) Comparisons between data clustering algorithms. *The International Arab Journal of Information Technology* **5**: 320-325.

Aizenstein HJ, Nebes RD, Saxton JA, Price JC, Mathis CA, Tsopelas ND, Ziolko SK, James JA, Snitz BE, Houck PR, Bi W, Cohen AD, Lopresti BJ, DeKosky ST, Halligan EM, Klunk WE. (2008) Frequent amyloid deposition without significant cognitive impairment among the elderly. *Archives of Neurology* **65**: 1509-1517.

Brown H, Prescott R. (2006) Applied mixed models in medicine. Wiley. Second edition. Page 224.

Celeux G, Lavergne C, Martin O. Mixture of linear mixed models application to repeated data clustering. ISSN 0249-6399 ISRN INRIA/RR-4566-FR+ENG.

Dempster AP, Laird NM, Rubin DB. (1977) Maximum likelihood from incomplete data via the EM-algorithm. *Journal of Royal Statistical Society B* **39**: 1-38.

Fitzmaurice GM, Laird NM, Ware JH. (2004) Applied longitudinal analysis. Wiley. Page 99.

Fraley C, Raftery AE. (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal* **41**: 578-588.

Gaffney S, Smyth P. (1999) Trajectory clustering with mixtures of regression models. Technical report No. 99-15 Department of Information and Computer Science, University of California, Irvine, CA, USA.

Grün B, Leisch F. (2008) FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software* **28**: 1-36.

Grün B, Scharl T, Leisch F. (2012) Modeling time course gene expression data with finite mixtures of linear additive models. *Bioinformatics* **28**: 222-228.

Horrocks J, van Den Heuvel MJ. (2009) Prediction of pregnancy: a joint model for longitudinal and binary data. *Bayesian Analysis* **4**: 523-538.

Hughes JP. (1999) Mixed effects models with censored data with application to HIV RNA levels. *Biometrics* **55**: 625-629.

Jacqmin-Gadda H, Thiebaut R, Chene G, Commenges D. (2000) Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics* **1**: 355-368.

Kale S, Yende S, Kong L, Perkins A, Kellum JA, Newman AB, Vallejo AN, Angus DC. (2010) The effects of age on inflammatory and coagulation fibrinolysis response in patients hospitalized for pneumonia. *PLoS One* **5**: e13852.

Kellum JA, Kong L, Fink MP, Weissfeld LA, Yealy DM, et al. (2007) Understanding the inflammatory cytokine response in pneumonia and sepsis: results of the Genetic and Inflammatory Markers of Sepsis (GenIMS) Study. *Arch Intern Med* **167**: 1655–1663.

Leisch F. (2004) Exploring the structure of mixture model components. In J˜Antoch (ed.), *Compstat 2004— Proceedings in Computational Statistics*: 1405-1412. Physica Verlag, Heidelberg, Germany. ISBN 3-7908-1554-3.

Lenk PJ, DeSarbo WS. (2000) Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika* **65**: 93-119.

Li E, Zhang D, Davidian M. (2004) Conditional estimation for generalized linear models when covariates are subject-specific parameters in a mixed model for longitudinal measurements. *Biometrics* **60**: 1-7.

Li E, Wang N, Wang NY. (2007) Joint models for a primary endpoint and multiple longitudinal covariate processes. *Biometrics* **63**: 1068-1078.

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

Lin H, McCulloch CE, Mayne ST. (2002) Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Statistics in Medicine* **21**: 2369-2382.

Luan Y, Li H. (2003) Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics* **19**: 474-482.

Ma P, Zhong W, Liu JS. (2009) Identifying differentially expressed genes in time course microarray data. *Statistics in Biosciences* **1**: 144-159.

Mellors JW, Munoz A, Giorgi JV, Margolick JB, Tassoni CJ, Gupta P, Kingsley LA, Todd JA, Saah AJ, DetelsR, Phair JP, Rinaldo Jr CR. (1997) Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Annals of Internal Medicine **126***: 946-954.

Nagin DS. (2009) Group-based modeling: an overview. Chapter 4 of Handbook on Crime and Deviance, Handbooks of Sociology and Social Research, Springer Science+Business Media, LLC pp. 59-73.

Proust-Lima C, LetenneurL, Jacqmin-GaddaH. (2007) A nonlinear latent class model for joint analysis of multivariate longitudinal data and a binary outcome. *Statistics in Medicine* **26**: 2229-2245.

Qin LX, Self SG. (2006) The clustering of regression models method with applications in gene expression data. *Biometrics* **62**: 526-533.

R Development Core Team. (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

SAS Institute Inc. 2008. SAS/STAT® 9.2 User's Guide The NLMIXED Procedure. Cary, NC: SAS Institute Inc.

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

Stefanski LA, Carroll RJ. (1987) Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika* **74**: 703-716.

Thiebaut R, Jacqmin-Gadda H. (2004) Mixed models for longitudinal left-censored repeated measures. *Computer Methods and Programs in Biomedicine* **74**: 255-260.

Thiebaut R, Jacqmin-Gadda H, Chene G, Leport C, Commenges D. (2002) Bivariatelinear mixed models using SAS pro MIXED. *Computer Methods and Programs in Biomedicine* **69**: 249-256.

Thiebaut R, Jacqmin-Gadda H, Babiker A, Commenges D, CASCADE Collaboration. (2005) Joint modeling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection. *Statistics in Medicine* **24**: 65-82.

Tseng GC, Wong WH. (2005) Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics* **61**: 10-16.

Tsiatis AA, Davidian M. (2004) Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14**: 809-834.

Wang CY, Huang Y. (2001) Functional methods for logistic regression on random-effect-coefficients for longitudinal measurements. *Statistics & Probability Letters* **53**: 347-356.

Wang CY, Wang N, Wang S. (2000) Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. *Biometrics* **56**: 487-495.

Wannemuehler KA, Lyles RH, Manatunga AK, Terrel ML, Marcus M. (2010) Likelihood-based methods for estimating the association between a health outcome and left-or interval-censored longitudinal exposure data. *Statistics in Medicine* **29**: 1661-1672.

Yeung KY, Medvedovic M, Bumgarner RE. (2003) Clustering Gene-expression Data with Repeated Measurements. *Genome Biology* **4**: R34.

Yuan M, Kendziorski C. (2005) A unified approach for simultaneous gene clustering and differential expression identification. *Statistics discussion paper* No. 2005-02.